2021 International Conference on Multimedia Analysis and Pattern Recognition (MAPR) | 978-1-6654-1910-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/MAPR53640.2021.9585242

Exploring Zero-shot Cross-lingual Aspect-based Sentiment Analysis using Pre-trained Multilingual Language Models

1rd Khoa Thi-Kim Phan
Department of Computer Science
University of Information Technology
Vietnam National University, Ho Chi Minh City
Vietnam
Email: 18520934@gm.uit.edu.vn

3nd Duong Ngoc Hao
Department of Mathematics and Physics
University of Information Technology
Vietnam National University, Ho Chi Minh City
Vietnam
Email: haodn@uit.edu.vn

2st Dang Van Thin
Multimedia Communications Laboratory (MMLab)
University of Information Technology
Vietnam National University, Ho Chi Minh City
Vietnam
Email: thindv@uit.edu.vn

4th Ngan Luu-Thuy Nguyen
Faculty of Information Science and Engineering
University of Information Technology
Vietnam National University, Ho Chi Minh City
Vietnam
Email: ngannlt@uit.edu.vn

***Abstract*—Aspect-based sentiment analysis (ABSA) has received much attention in the Natural Language Processing research community. Most of the proposed methods are conducted exclusively in English and high-resources languages. Leveraging resources available from English and transferring to low-resources languages seems to be an immediate solution. In this paper, we investigate the performance of zero-shot cross-lingual transfer learning based on pre-trained multilingual models (mBERT and XLM-R) for two main sub-tasks in the ABSA problem: Aspect Category Detection and Opinion Target Expression. We experiment on the benchmark data sets of six languages as English, Russian, Dutch, Spanish, Turkish, and French. The experimental results demonstrated that using the XLM-R model can yield relatively acceptable results for the zero-shot cross-lingual scenario.**

***Index Terms*—zero-shot cross-lingual, aspect based sentiment analysis, multilingual language model, aspect category detection, opinion target expression.**

I. Introduction

Over the past decade, Sentiment Analysis(SA) is one of the most popular tasks in the Natural language processing (NLP) field due to the evolution of the Internet, particularly the increasing amount of user opinion content. It aims to identify and extract user opinions[1], often in categories like positive, neutral, and negative. However, sentiment analysis may not be enough in response to all demands in the real world if the given text has more than one topic or aspect. For example, we have a sentence, "This food is great, but the waitress has a bad attitude." In this example, we get two sentiment polarities toward two aspects: "food" is positive and "service" is negative. Therefore, a more detailed version can identify the sentiment polarity of given aspects in an input sentence, known as aspect-based sentiment analysis(ABSA).

To solve this task, supervised learning algorithms are a strong candidate because of their effectiveness [2]–[4]. Nevertheless, the supervised learning-based approach requires large, domain-specific datasets and manual training data [5], which is expensive and time-consuming. If we apply this approach to a new language, it is a challenge, especially low-resourced languages [6]. Comparable to [7], in this paper, we try to utilize available annotated datasets called source language and predict a target language, namely Zero-shot cross-lingual transfer to deal with ABSA task. Specifically, we present an approach based on pre-trained multilingual language models mBERT [8] and XLM-R [9] for zero-shot cross-lingual ABSA evaluation.

Our paper makes the following contributions:

- We investigate the benefit of using pre-trained multilingual models (mBERT and XLM-R) to evaluate zero-shot cross-lingual learning for two main sub-tasks in the ABSA problem: Aspect Category Detection and Opinion Target Expression.
- We experiment on the benchmark datasets of six languages, including English, Russian, Dutch, Spanish, Turkish, and French. In addition, we point out which model is better for each of two tasks and each of benchmark datasets of six languages.

The rest of the paper is structured as follows. Section 2 is an overview of related work about aspect-based sentiment analysis. Section 3 presents the model which is used for experiments in this paper. Then, we present the whole experiments, including datasets, model settings, and our results in Section 4. Section 5 presents the conclusion and gives some future

TABLE I: The input and the output of two tasks in the ABSA problem.

Input: Sometimes I get good food and ok service.	
Task	**Output**
Opinion Target Expression	food; service
Aspect Category Detection	Food#Quality; Service#General

works.

II. Related Work

In recent years, Aspect-based Sentiment Analysis(ABSA) has attracted the most attention due to its thorough analysis and real world application. It was first introduced as Task 4 in SemEval-2014 [10], and appeared again in SemEval-2015 Task 12 [11] and SemEval-2016 Task 5 [12].

A number of state-of-the-art ABSA studies in English require sophisticated NLP tools or hand-crafted sentiment lexicons [13]. However, nowadays, with the development of Language Representation Learning in NLP and the challenge to build large-scale labeled datasets for most NLP tasks due to the extremely expensive annotation costs, especially for syntax and semantically related tasks [14], pre-trained models (PTMs) has proven as being beneficial for downstream NLP tasks. Besides, learning good word embeddings, PTMs focus on learning contextual word embeddings, such as ELMo [15], OpenAI GPT [16], BERT [8], and its variants, etc.

Especially, BERT and its variants have achieved most of state-of-the-art in many NLP tasks. Therefore, several studies have been conducted on ABSA. [17] proposed the models with the aspect tokens and text connected and used as the input of BERT. [18] proposed a method to solve ABSA as a sentence-pair classification task using BERT by constructing an auxiliary sentence. [19] analyzed the pre-trained hidden representations learned from reviews on BERT for tasks in aspect-based sentiment analysis (ABSA). [20] explored the potential of utilizing BERT intermediate layers to enhance the performance of fine-tuning of BERT.

In addition, the ABSA studies are not only for English, but it is also for other languages. [21] presented an ABSA using French PTM like multilingual BERT (mBERT) [22], CamemBERT [23] and FlauBERT [24]. Then, they fine-tuned models by three fine-tuning methods: Fully-Connected, Sentences Pair Classification, and Attention Encoder Network. [25] employed ParsBERT pre-trained model along with natural language inference auxiliary sentence (NLI-M) to improve the ABSA on the Persian Pars-ABSA dataset.

III. Approach

This section presents two models based on fine-tuning pre-trained multilingual language models to solve two main sub-tasks in the ABSA problem. In addition, we also present briefly about Zero-shot Cross-Lingual Scheme. The example of input and output corresponding to two sub-tasks is displayed in Table I.

A. Opinion Target Expression

The purpose of this task aims to extract the phrases which indicate the target aspect in the input. As shown in the Table I, the input of this task is a user's review, and the output is two words - "food" and "service" mentioned directly in the review. To tackle this task, a common approach is to convert it to a sequence tagging problem using the well-known IOB scheme. According to this scheme, each word in the input is assigned one of three labels: I, O, B where indicate the Begin, Inside, and Outside of a target expression. An example of this approach can be seen below:

Sometimes	*I*	*get*	*good*	*food*	*and*	*ok*	*service*	.
O	O	O	O	B_{TAR}	O	O	B_{TAR}	O

Based on this approach, we can solve this task by using sequence tagging models. In this paper, we use a fine-tuning model based on transformer architecture for the sequence tagging problem. This model adds a linear layer as a token-level classifier that takes the last hidden state of the sequence of the BERT architectures. The model takes a sequence of *n* word: $X = \{w_1, w_2, ..., w_n\}$ as input and predicts an output sequence of IOB tags: $Y = \{O, O, ..., B_{TAR}\}$. We use the BERT model to extract the last hidden state of tokens as the representation and put them into the output layer to predict the tag for tokens.

B. Aspect Category Detection

The purpose of this task is to extract the aspect category mentioned in the sentence. As can be seen in the Table I, the input is the user's review; the output is two categories as Food#Quality and Service#General. The number of aspect categories is pre-defined by the domain of the dataset. For the restaurant domain, there are 12 different aspect categories used for the experiments. We can see that each comment can contain one or more categories. Therefore, this task can be treated as a multi-label classification problem. It means that the number of nodes in the output layer is the number of aspect categories. To solve this task, the output can be represented as a binary vector where each index is 0 or 1, where indicate the corresponding aspect categories. Therefore, we use the simple approach based on the pre-trained language models for this task. We extract the representation of [CLS] token in the last layer as the input representation. This representation is fed into the fully connected layer with Sigmoid activation, which predicts a probability of class for the label a value between 0 and 1. Because the probability of each value in the output vector is independent, we need a threshold to determine whether the label is assigned to the input. In our work, we set the value of the threshold as 0.5. Finally, we use the Binary Cross-Entropy as the loss function.

C. Zero-shot Cross-Lingual Scheme

The definition of zero-shot cross-Lingual transfer is aimed to build models for evaluating the testing data in a target language by reusing knowledge acquired from the learned models on training data in a source language[26]–[28]. Therefore, we

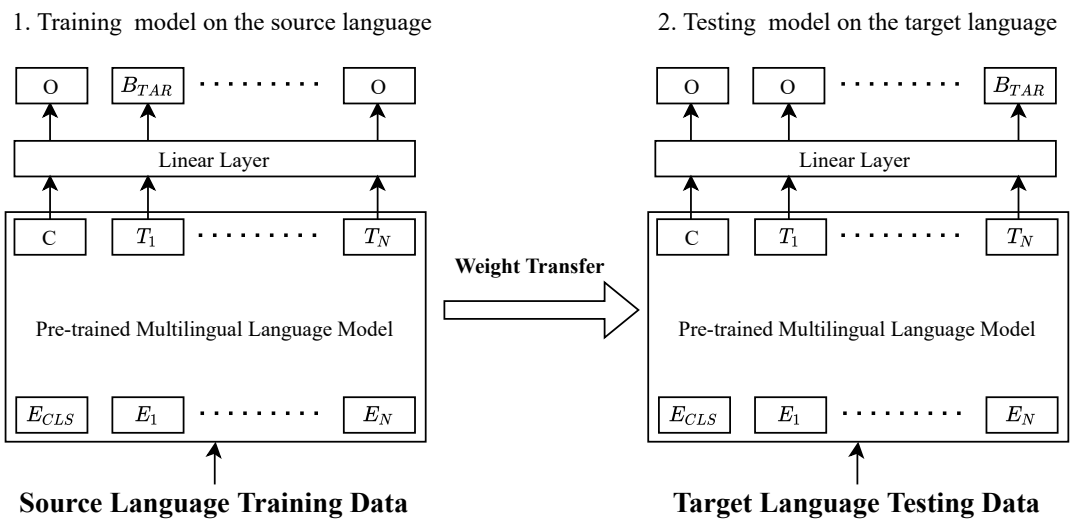

Fig. 1: Zero-shot cross-lingual evaluation for the Aspect Term Extraction task.

TABLE II: Statistics of the SemEval workshop 2016 Task 5 for the restaurant domain.

Dataset	Language	#Sent	#Targets	#Categories
Train	English	2000	1880	2507
	French	1733	1770	2530
	Dutch	1722	1283	1860
	Spanish	2070	1937	2720
	Turkish	1232	1385	1535
	Russian	3655	3159	4089
Test	English	676	650	859
	French	696	718	954
	Dutch	575	394	613
	Spanish	881	731	1072
	Turkish	144	159	159
	Russian	1209	972	1300

present a simple approach based on pre-trained multilingual language models for zero-shot cross-lingual evaluation as illustrated in Figure 1. In this paper, we investigate two learning settings: (1) *zero-shot cross-lingual* - i.e., training on source language and testing on target languages, for example, we fine-tune the pre-trained XLM-R model on the English training set and evaluate this model on the testing set of other languages such as Russian, Spanish, etc.; (2) *monolingual*, i.e., training and testing on the same language. In our work, we employ two popular pre-trained language models as mBERT [8] and XLM-R model [9]. The mBERT model is trained on the Wikipedia data of 104 languages, while the XLM-R model is trained on the CommonCrawl data in 100 languages. Both models include the languages in our study.

IV. Experiments

A. Datasets

We use a collection of datasets on various languages for the restaurant domain, which are presented at SemEval 2016 workshop Task 5 [29]. This collection includes the training set and testing set for the six languages: English, French, Russian, Spanish, Dutch, and Turkish. Table II shows the statistic of the experimental datasets.

B. Experimental Settings

In all our experiments, we report the F_1-score and micro F_1-score for the Opinion Target Expression and Aspect Category Detection task as in the original SemEval task [29], respectively.

As described in Section III, our models rely on pre-trained language models such as mBERT and XLM-R model. We use two base models downloaded from the Hugging Face library [30]. The network's parameters are optimized using the AdamW with warm-up steps for 10% of the training data and a learning rate of 5e-5. For the OTE task, we train two models up to 1500 steps. For the ACD task, we set the number of epochs as 50. For the pre-processing steps, we do not apply any step for the used datasets.

C. Result and Discussion

In this section, we present our evaluation for zero-shot cross-lingual learning based on two pre-trained multilingual models. We first examine the performance of two pre-trained language models trained and tested on a target language. Then, we analyze the zero-shot cross-lingual performance on the language pairs. Figure 2 and Figure 3 show the results of pair languages corresponding to the ACD an OTE task, respectively. The main diagonal of each matrix represents the results of models on a specific language.

In general, fine-tuning the XLM-R model achieved better results than the mBERT model for all languages for the ACD task and some languages for the OTE task in the monolingual setting. As shown in Figure 2, the performance of the XLM-R model is higher than the mBERT model, about 4.33% to 9.18% in the range of languages. In contrast, the mBERT model has only achieved better results in French and Dutch language for the Opinion Target Expression task. Specifically, the mBERT

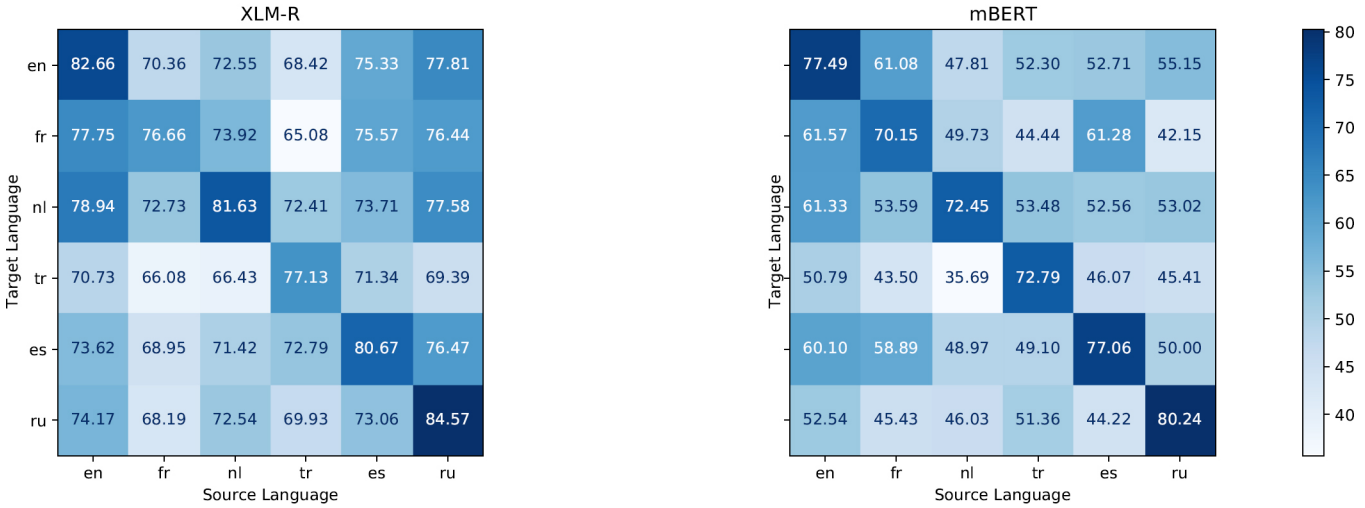

Fig. 2: Evaluation of zero-shot cross-lingual approach based on the two pre-trained language models for the Aspect Category Detection task.

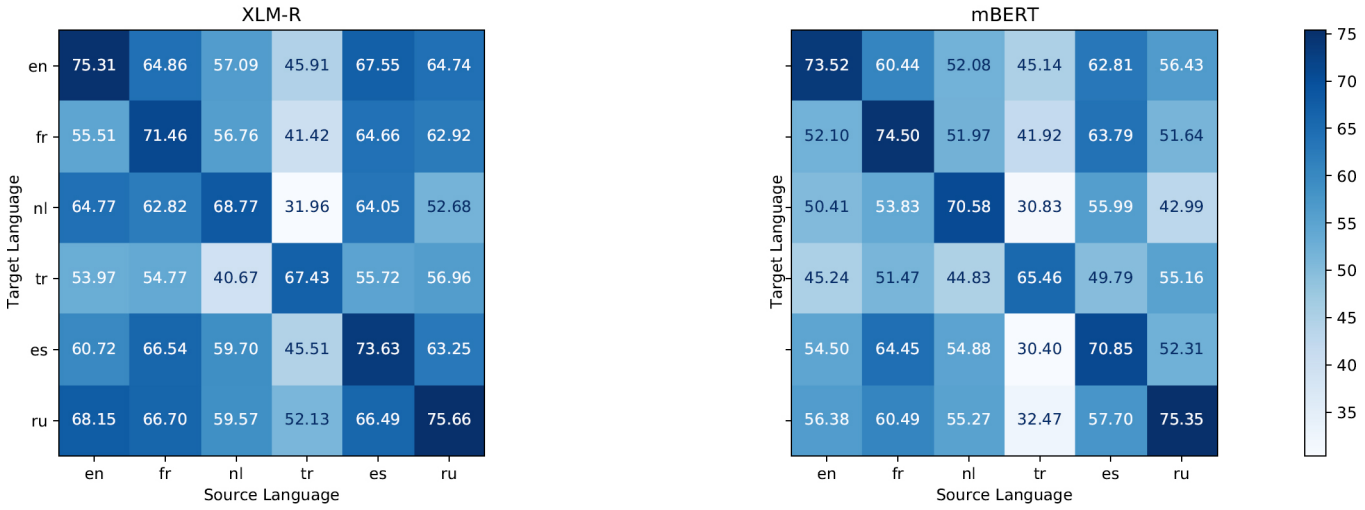

Fig. 3: Evaluation of zero-shot cross-lingual approach based on the two pre-trained language models for the Opinion Target Expression task.

model reached of F1-score of 74.50% and 70.58%, while the XLM-R model gained an F1-score of 71.46% and 68.77% corresponding to the French and Dutch language, respectively. There are many reasons to interpret our experimental results. The main reasons might be the XLM-R model using more pre-training data and various domains than the mBERT model. A previous work [31] demonstrated that using more data can improve the quality of the pre-trained language models. In detail, the XLM-R model is trained on the combination of Common Crawl with Wikipedia data, while mBERT is only trained on Wikipedia data. XLM-R achieved high micro-F1 scores in monolingual settings for detecting the ACD task in English, Dutch, Spanish, and Russian (82.66 for EN $\rightarrow$ EN; 81.63 for NL $\rightarrow$ NL; 80.67 for ES $\rightarrow$ ES; 84.57 for RU $\rightarrow$ RU). For the OTE task, the mBERT model achieved better results in the French and Dutch languages. This can demonstrate that the contextual representation of each token of the mBERT model for French and Dutch language is better than the XLM-R model for the token classification task. Our results are similar to the recently presented findings [32].

For the zero-shot cross-lingual experiments, the performance of the XLM-R model is also better than the mBERT model across language pairs for two tasks. Besides, we observe a drop of the micro F1-score when considering the zero-shot cross-lingual settings for both tasks; however, this finding depends on the target language. In detail, for the ACD task, the percentage reduction of XLM-R model on the zero-shot cross-lingual comparing to monolingual setting as follows: -3.72% for EN, -3.93% for FR, -7.71% for NL, -4.34% for TU, -5.10% for ES, and -6.76% for RU. For the OTE task,

TABLE III: Zero-shot results with the monolingual results of two models for two tasks. The column $best \rightarrow target$ shows the best performance of zero-shot cross-lingual from Table 2 and 3. The column $target \rightarrow target$ presents the results of monolingual scores.

Task	Method	best → target						target → target					
		en	fr	nl	tu	es	ru	en	fr	nl	tu	es	ru
OTE	XLM-R	68.15(ru)	66.70(ru)	59.70(es)	52.13(ru)	67.55(en)	64.74(en)	75.31	71.46	68.77	67.43	73.63	75.66
	mBERT	56.38(ru)	64.45(es)	55.27(ru)	45.14(en)	63.79(fr)	56.43(en)	73.52	74.50	70.58	65.46	70.85	75.35
ACD	XLM-R	78.94(nl)	72.73(nl)	73.92(fr)	72.79(es)	75.57(fr)	77.81 (en)	82.66	76.66	81.63	77.13	80.67	84.57
	mBERT	61.57(fr)	61.08(en)	49.73(fr)	53.48(nl)	52.71(en)	55.15(en)	77.49	70.15	72.45	72.79	77.06	80.24

the performance of zero-shot cross-lingual reduce more than monolingual setting: -7.16% for EN, -4.76% for FR, -9.07% for NL, -15.3% for TU, -6.08% for ES, and -7.92% for RU. Looking at Table III, we can observe the best language pairs for the target language. For example, fine-tuning the XLM-R model on the Russian training set achieved the best performance on the English testing set with an F1-score of 68.15%. For the OTE task, the Russian language is the best source language in terms of zero-shot cross-lingual for the target language, followed by the English source language. However, the best language pairs for the target language are different for the ACD task. For instance, fine-tuning XLM-R languages from the source Dutch language achieved the best performance on the English test language. In general, we can see that training the Turkish language seems to give lower performance for other target languages, while the Russian language is the best source language for the target language for both tasks. This can be explained because of the number of training samples in Turkish language (see Table II).

V. Conclusion And Future Work

In this work, we presented an investigation for zero-shot cross-lingual on two main sub-tasks of aspect-based sentiment analysis. In addition, we also compared the performance of two models on a specific language that trained and tested on the target language data. Our models are entirely based on two famous pre-trained multilingual language models (mBERT and XLM-R), which allow us to evaluate six languages.

For the Aspect Category Detection task, we have also found that fine-tuning the XLM-R model yields better performance than the mBERT model for six languages for the monolingual setting. Besides, the XLM-R also achieved better results than the mBERT model on some languages for the Opinion Target Expression task. For the zero-shot cross-lingual scenario, using the Russian and English language as a source language was able to reach the best F1-score in some languages for the OTE task. While the sequence classification task as ACD, the language pairs are different based on language similarity. For future work, we plan to expand this evaluation to other low-source languages and investigate a solution based on the combination of deep learning and transformer models.

Acknowledgment

We thank the anonymous reviewers for their valuable comment for this manuscript. This research is funded by the University of Information Technology - Vietnam National University Ho Chi Minh City under grant number D1-2021-25.

References

[1] E. Cambria, S. Poria, A. Gelbukh, and M. Thelwall, "Sentiment analysis is a big suitcase," *IEEE Intelligent Systems*, vol. 32, no. 6, pp. 74–80, 2017.

[2] N. U. Pannala, C. P. Nawarathna, J. Jayakody, L. Rupasinghe, and K. Krishnadeva, "Supervised learning based approach to aspect based sentiment analysis," in *2016 IEEE International Conference on Computer and Information Technology (CIT)*, IEEE, 2016, pp. 662–666.

[3] J. Macháček, "Butknot at semeval-2016 task 5: Supervised machine learning with term substitution approach in aspect category detection," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 2016, pp. 301–305.

[4] T. Álvarez-López, J. Juncal-Martínez, M. F. Gavilanes, E. Costa-Montenegro, and F. González-Castaño, "Gti at semeval-2016 task 5: Svm and crf for aspect detection and unsupervised aspect-based sentiment analysis," in **SEMEVAL*, 2016.

[5] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004, pp. 168–177.

[6] N. T. T. Thuy, N. X. Bach, and T. M. Phuong, "Cross-language aspect extraction for opinion mining," in *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, IEEE, 2018, pp. 67–72.

[7] S. Jebbara and P. Cimiano, "Zero-shot cross-lingual opinion target extraction," *arXiv preprint arXiv:1904.09122*, 2019.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 NAACL: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[9] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Jul. 2020, pp. 8440–8451.

[10] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar, "SemEval-2014 task 4: Aspect based sentiment analysis," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland: Association for Computational Linguistics, Aug. 2014, pp. 27–35.

[11] M. Pontiki, D. Galanis, H. Papageorgiou, S. Manandhar, and I. Androutsopoulos, "SemEval-2015 task 12: Aspect based sentiment analysis," in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, Denver, Colorado: Association for Computational Linguistics, Jun. 2015, pp. 486–495.

[12] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. AL-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, O. De Clercq, V. Hoste, M. Apidianaki, X. Tannier, N. Loukachevitch, E. Kotelnikov, N. Bel, S. M. Jiménez-Zafra, and G. Eryiğit, "SemEval-2016 task 5: Aspect based sentiment analysis," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 19–30.

[13] J. Barnes, P. Lambert, and T. Badia, "Exploring distributional representations and machine translation for aspect-based cross-lingual sentiment classification.," in *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*, 2016, pp. 1613–1623.

[14] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "Pre-trained models for natural language processing: A survey," *Science China Technological Sciences*, pp. 1–26, 2020.

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

[16] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018.

[17] M. Hoang, O. A. Bihorac, and J. Rouces, "Aspect-based sentiment analysis using bert," in *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, 2019, pp. 187–196.

[18] C. Sun, L. Huang, and X. Qiu, "Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence," *arXiv preprint arXiv:1903.09588*, 2019.

[19] H. Xu, L. Shu, P. S. Yu, and B. Liu, "Understanding pre-trained bert for aspect-based sentiment analysis," *arXiv preprint arXiv:2011.00169*, 2020.

[20] Y. Song, J. Wang, Z. Liang, Z. Liu, and T. Jiang, "Utilizing bert intermediate layers for aspect based sentiment analysis and natural language inference," *arXiv preprint arXiv:2002.04815*, 2020.

[21] A. Essebbar, B. Kane, O. Guinaudeau, V. Chiesa, I. Quénel, and S. Chau, "Aspect based sentiment analysis using french pre-trained models.," in *ICAART (1)*, 2021, pp. 519–525.

[22] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual bert?" *arXiv preprint arXiv:1906.01502*, 2019.

[23] L. Martin, B. Muller, P. J. O. Suárez, Y. Dupont, L. Romary, É. V. de La Clergerie, D. Seddah, and B. Sagot, "Camembert: A tasty french language model," *arXiv preprint arXiv:1911.03894*, 2019.

[24] H. Le, L. Vial, J. Frej, V. Segonne, M. Coavoux, B. Lecouteux, A. Allauzen, B. Crabbé, L. Besacier, and D. Schwab, "Flaubert: Unsupervised language model pre-training for french," *arXiv preprint arXiv:1912.05372*, 2019.

[25] H. Jafarian, A. H. Taghavi, A. Javaheri, and R. Rawassizadeh, "Exploiting bert to improve aspect-based sentiment analysis performance on persian language," in *2021 7th International Conference on Web Research (ICWR)*, IEEE, 2021, pp. 5–8.

[26] A. Eriguchi, M. Johnson, O. Firat, H. Kazawa, and W. Macherey, "Zero-shot cross-lingual classification using multilingual neural machine translation," *arXiv preprint arXiv:1809.04686*, 2018.

[27] B. Muller, Y. Elazar, B. Sagot, and D. Seddah, "First align, then predict: Understanding the cross-lingual ability of multilingual bert," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2021, pp. 2214–2231.

[28] K.-H. Huang, W. U. Ahmad, N. Peng, and K.-W. Chang, "Improving zero-shot cross-lingual transfer learning via robust training," *arXiv preprint arXiv:2104.08645*, 2021.

[29] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. Al-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, O. De Clercq, *et al.*, "Semeval-2016 task 5: Aspect based sentiment analysis," in *International workshop on semantic evaluation*, 2016, pp. 19–30.

[30] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.

[31] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[32] H. You, X. Zhu, and S. Stymne, "Uppsala nlp at semeval-2021 task 2: Multilingual language models for fine-tuning and feature extraction in word-in-context disambiguation," *arXiv preprint arXiv:2104.03767*, 2021.