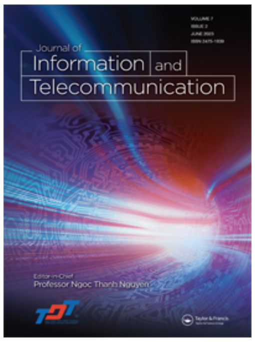

Taylor & Francis
Taylor & Francis Group

Journal of Information and Telecommunication

ISSN: 2475-1839 (Print) 2475-1847 (Online) Journal homepage: www.tandfonline.com/journals/tjit20

# Exploring zero-shot and joint training cross-lingual strategies for aspect-based sentiment analysis based on contextualized multilingual language models

Dang Van Thin, Hung Quoc Ngo, Duong Ngoc Hao & Ngan Luu-Thuy Nguyen

JOURNAL OF INFORMATION AND TELECOMMUNICATION
2023, VOL. 7, NO. 2, 121–143
https://doi.org/10.1080/24751839.2023.2173843

OPEN ACCESS Check for updates

# Exploring zero-shot and joint training cross-lingual strategies for aspect-based sentiment analysis based on contextualized multilingual language models

Dang Van Thin [a,b], Hung Quoc Ngo [c], Duong Ngoc Hao [a,b] and Ngan Luu-Thuy Nguyen [a,b]

[a]Multimedia Communications Laboratory, University of Information Technology, Ho Chi Minh city, Vietnam; [b]Vietnam National University, Ho Chi Minh city, Vietnam; [c]School of Business Technology, Retail, and Supply Chain, Technological University Dublin, Dublin, Ireland

**ABSTRACT**
Aspect-based sentiment analysis (ABSA) has attracted many researchers' attention in recent years. However, the lack of benchmark datasets for specific languages is a common challenge because of the prohibitive cost of manual annotation. The zero-shot cross-lingual strategy can be applied to solve this gap in research. Moreover, previous works mainly focus on improving the performance of supervised ABSA with pre-trained languages. Therefore, there are few to no systematic comparisons of the benefits of multilingual models in zero-shot and joint training cross-lingual for the ABSA task. In this paper, we focus on the zero-shot and joint training cross-lingual transfer task for the ABSA. We fine-tune the latest pre-trained multilingual language models on the source language, and then it is directly predicted in the target language. For the joint learning scenario, the models are trained on the combination of multiple source languages. Our experimental results show that (1) fine-tuning multilingual models achieve promising performances in the zero-shot cross-lingual scenario; (2) fine-tuning models on the combination training data of multiple source languages outperforms monolingual data in the joint training scenario. Furthermore, the experimental results indicated that choosing other languages instead of English as the source language can give promising results in the low-resource languages scenario.

**ARTICLE HISTORY**
Received 14 September 2022
Accepted 29 November 2022

**KEYWORDS**
Aspect-based sentiment analysis; zero-shot cross-lingual; multilingual models; joint learning

## 1. Introduction

Nowadays, the development of social networks and e-commerce helps users share and quickly consult feedback about products and services of business organizations. Customers tend to refer to comments before making decisions. In addition, users' comments are also valuable resource for business organizations to analyse, develop, and improve their products and services to provide the best customer experience. Unfortunately,

**CONTACT** Dang Van Thin thindv@uit.edu.vn

manual processing by human annotation is not possible for massive comments. Hence, the Opinion Mining task has attracted much attention from researchers worldwide and business organizations in the field of Natural Language Processing (NLP). Most of the research has recently focused on solving this task at the aspect level, called Aspect-based Sentiment Analysis. Therefore, more insight information can be extracted from the tremendous comments automatically. For example, given a review for the restaurant domain as '*This place is great, but the food is not delicious*'. There are two aspect categories (Restaurant#General and Food#Quality) in this sentence; but the sentiment polarity of categories is contradictory (positive for Restaurant#General, negative for Food#Quality). It is obvious that the polarity for two aspect categories is different. Analysing the comments on positive or negative categories has been able to improve the service and attract new customers.

There are currently more than 7000 languages worldwide Joshi et al. (2020a); however, most recent research focuses only on resource-rich languages such as English, Chinese, Arabic, etc. These languages have a lot of abundant and diverse annotated resources for various NLP tasks and tools. Moreover, building a new dataset requires more resources and costs to manually annotate for a specific language. Therefore, the lack of annotated datasets for low-resource languages has become a challenge for researchers in the NLP field. Recently, the great work of Hedderich et al. (2021) presented an overview of different approaches to improving the performance of low-resource languages, including data enhancement, multilingual language models, etc. Following the success of multilingual BERT Devlin et al. (2019), there are many pre-trained transformer multilingual models such as XLM-R (Conneau et al., 2020), InfoXLM (Chi et al., 2021a), XLM-Align (Chi et al., 2021b), which are beneficial for low-resource languages, with the task-specific annotated data being scarce. There are several recent studies that take advantage of exist pre-trained multilingual language model to improve the performance of system on different tasks such as sentiment analysis (Kumar & Albuquerque, 2021; Pei et al., 2022; Sarkar et al., 2019; Sultan et al., 2020), Named Entity Recognition (Arkhipov et al., 2019; Pires et al., 2019; Sharma et al., 2022b), Hate Speech Detection (Sharma et al., 2022a), and other tasks (Bhatnagar et al., 2022; Sun et al., 2022). In addition, pre-trained multilingual language models have been applied in zero-shot (Artetxe & Schwenk, 2019; Keung et al., 2020; Kim et al., 2021; Lauscher et al., 2020a; Nooralahzadeh et al., 2020; Pamungkas et al., 2021; Phan et al., 2021) cross-lingual for various NLP tasks. However, most of these studies only compared the performance of mBERT to XLM-R models as well as used English as a source language. There is no fair comparison when the amount of training data is different between languages in the zero-shot learning scenario.

There is no study on exploring the state-of-the-art (SOTA) pretrained multilingual language models to ABSA in both zero-shot and joint training cross-lingual scenarios. Therefore, the major objective of this study is to investigate the effectiveness of fine-tuning multilingual contextualized language models in both zero-shot and joint training cross-lingual scenarios on two main tasks of ABSA: Aspect Category Detection (ACD) and Category-Sentiment Classification (CSC). In this work, we make important contributions to the ABSA task, including:

- Firstly, we explore the power of zero-shot transfer learning for five languages in the context of lacking labelled training data in the target resource-poor language.

- Secondly, we conduct experiments to answer a research question: 'Why do not we use another language as the source language instead of using English for ABSA tasks in the zero-shot scenario?'. Because many previous studies have trained models on English data and tested them on non-English languages (Keung et al., 2020; Lin et al., 2019) in the zero-shot setting.
- Thirdly, we evaluate the system's performance in a joint training strategy by mixing training data of the source language and target language and combining multiple source languages.
- Finally, many of the latest multilingual language models as InfoXLM, XLM-Align have not been explored before on zero-shot and joint learning cross-lingual task, especially resource-low languages. In this paper, we also investigate several latest pre-trained multilingual transformer models for two tasks in the ABSA problem.

The remainder of this paper is structured as follows: Section 2 presents a survey of previous studies on ABSA task, zero-shot and joint learning research. Section 3 describes the methodology using different pre-trained language models in zero-shot and joint learning scenarios. Section 4 presents the experimental results. Our conclusions are found in the final section.

## 2. Related work

This section consists of three sub-sections covering the associated studies for the ABSA problem, zero-shot and joint training cross-lingual in the NLP field. The purpose of this paper is to explore the performance of zero-shot and joint training cross-lingual for ABSA tasks. Therefore, we survey the most recent work concerning the ABSA in Section 2.1, zero-shot learning in Section 2.2 and joint learning in Section 2.3.

### *2.1. Aspect-based sentiment analysis*

In recent years, the power of contextual language models has increased the performance in system to the field of the ABSA. First, there are many datasets published for the research community at shared-task SemEval 2014 (Pontiki et al., 2014), SemEval 2015 (Pontiki et al., 2015), SemEval 2016 (Pontiki et al., 2016). The shared-task SemEval provided several datasets from various domains in languages such as English, Chinese, Dutch, etc. These datasets are very popular and are benchmark datasets for many ABSA tasks. Recently, the pre-trained BERT (Devlin et al., 2019) language models have shown their effectiveness in various tasks in ABSA problems.

Sun et al. (2019) presented four new methods based on fine-tuning BERT with an auxiliary sentence for T(ABSA) problem. They transformed this problem into a sentence-pair classification task and fine-tuned the pre-trained BERT model. Their experimental results demonstrated the advantages of sentence pair classification based on the BERT model for the ABSA task, however, their models take a lot of computation resources and time for training. Hoang et al. (2019) presented an overview of fine-tuning the pre-trained BERT model to address the out-of-domain ABSA problem at both levels of datasets by using the sentence pair classification approach. However, the authors just conducted the experiments on the English language instead of other languages in the SemEval datasets. Li et al. (2019) presented an end-to-end neural-based on BERT architecture for the aspect term with corresponding sentiment polarity. They formulated two tasks as a sequence labelling problem and used the pre-trained BERT embedding as the embedding layer combined with several different layers (linear layer, recurrent network, self-attention, and conditional random fields layer) on top of BERT. Their experimental results showed that the BERT-based model is a powerful architecture to improve the performance of aspect-based sentiment analysis problems.

On the other hand, Xu et al. (2019) proposed a BERT-based post-training model for the OTE task and Review reading comprehension (RRC) to enhance the domain-awareness. Due to the difference between the training corpus of BERT and the review corpus, this novel post-training to adapt BERT using two unsupervised objectives on the task-specific corpus to learn the domain-awareness contextualized representations. Rietzler et al. (2020) analysed the behaviour of domain-specific and cross-domain post-training techniques based on BERT language modelling for the Aspect-Target Sentiment Classification task. The experimental results indicated that domain-specific language model fine-tuning produce the state-of-the-art performance. Unfortunately, we have to provide enough a domain-specific corpora and resources to train the BERT model for a specific domain. This might may not be feasible for low-resource language and computationally-insufficient studies. Subsequently, Karimi et al. (2021) showed that using adversarial training with domain-specific post-trained BERT could further improve ABSA performance. In addition, they also investigated the number of training epochs and dropout values that can significantly affect on model's performance. Song et al. (2020) investigated the potential of BERT intermediate layers to improve the performance of BERT fine-tuning using the LSTM pooling or the attention mechanism. The experimental results demonstrated the effectiveness of the proposed approach to the ABSA problem. Wan et al. (2020) proposed a novel architecture that relied on the BERT language model to address the limitation of implicit terms in the review. Their model is able to capture the dependence of sentiments on both term expressions and aspect categories in the sequence by jointly learning. We can see that most of the above research focus on high-resource language such as English by leverage the available monolingual language models to improve the performance on supervised learning in ABSA problem.

## 2.2. *Zero-shot cross-lingual*

The development of deep learning architectures has achieved significant success in many areas; however, it requires a sufficient amount of labelled training data (Wang et al., 2019). Furthermore, labelling processing is time-consuming and expensive for several tasks; therefore, zero-shot learning methods (Larochelle et al., 2008) have been studied on various topics in the field of NLP, especially for languages without resources to train supervised models.

Jebbara and Cimiano (2019) addressed the lack of available annotated data for specific languages by applying a zero-shot cross-lingual approach for the opinion target expressions task. To do that, the author used the alignment of embeddings to calculate the cross-lingual representation of two languages based on the FastText embedding (Bojanowski et al., 2017). Jebbara and Cimiano (2019) presented the experiments using the Convolutional Neural Network as a baseline model and demonstrated the effectiveness of zero-shot learning. However, in this work, the authors ignore the influence of the data size of the target language. Lauscher et al. (2020b) presented extensive experiments for zero-shot cross-lingual transfer using multilingual pre-trained language models (mBERT, XLM-R) on different NLP tasks. The authors analysed the conditions and factors that affect the performance of cross-lingual transfer, such as the linguistic similarity and size of pre-training data. van der Heijden et al. (2020) presented a comprehensive comparison of a multilingual word and sentence representation for Named Entity Recognition and Part-of-Speech task in zero-shot learning settings. The results showed that pre-trained multilingual BERT outperformed other supervised models. However, the authors compared the mBERT model to the XLM transformer model (Conneau & Lample, 2019) instead of XLM-R.

Recently, Pamungkas et al. (2021) investigated the zero-shot learning approach for hate speech detection based on knowledge from resource-rich language. However, the author only transferred knowledge from English; therefore, the effectiveness of using knowledge of other languages has not been studied in the article. Kim et al. (2021) presented a parallel-labelled cross-lingual named entity recognition in English and Korean to develop a zero-shot learning model. They fine-tuned the mBERT in English and transferred the trained model to Korean, and compared it to the embedding and annotation projection approach. The experimental results showed that the order of words in the target language is important in cross-lingual learning. Kumar and Albuquerque (2021) applied the power of the XLM-R model to transfer knowledge from the English to Hindi dataset for the sentiment analysis dataset. Unfortunately, the author just compared the performance of the XLM-R large model to deep learning approaches; it is difficult to conclude that the XLM-R model is suitable for zero-shot scenario. Phan et al. (2021) presented a study on zero-shot cross-lingual learning on pre-trained multilingual models (mBERT and XLM-R) for two sub-tasks in ABSA problem. Unfortunately, the authors did not pay attention to the number of training samples among languages. This leads to unfair comparisons between models and languages.

### *2.3. Joint training*

The joint training scenario is an idea of training one model on multiple languages because many languages share common features such as morphological, phonological, and syntactic phenomena (Ammar et al., 2016; Bender, 2011; Mulcaire et al., 2018). As a result, training in multiple languages can improve the performance of models in related languages. Ammar et al. (2016) found that the training model on multilingual treebanks of multiple languages outperformed the monolingual training data for parsing tasks. However, the authors employed the traditional deep learning model (LSTM) combined with static multilingual word embedding instead of contextual word representation. Mulcaire et al. (2018) also applied this idea by combining training data across languages for semantic role-labelling tasks. The experimental results showed that joint learning could achieve better performance than monolingual data. Aharoni et al. (2019) presented extensive experiments in multilingual neural machine translation by training multilingual languages in a single model. This demonstrated that multilingual joint learning has been shown to be beneficial in various NLP tasks. The authors employed the XLM-R language model as the baselines. Recently, the development of multilingual pre-trained language models brings a lot of benefits to low-resource languages. With plenty of pre-trained language models and languages, how to choose them to improve the performance of a specific language is an interesting problem.

## 3. Cross-lingual framework

Figures 1 and 2 provide a comprehensive view of the zero-shot and joint learning cross-lingual learning on our experiments. We conduct two strategies on two main ABSA tasks, including Aspect Category Detection and Category-Sentiment Classification tasks. First, we present the problem formulation of two experimental tasks. Second, we summarize the methodology we used to build based models corresponding to two tasks. Third, we present the detail of zero-shot cross-lingual transfer learning approach based on the based models of five languages. Finally, we explain the cross-lingual joint training approach in this paper. As shown in Figure 1, the model is trained with training data of one source language (e.g. Spanish) based on a multilingual pretrained language model. Then the trained model is tested on target languages (e.g. English, French) in a zero-shot manner. While in the joint learning cross-lingual, the model is trained on the combined data of two source languages (e.g. Spanish and French) as in Figure 2.

### 3.1. Problem formulation

- *Aspect Category Detection*: The purpose of this task is to identify the pre-defined list of the entity E and attribute A pairs towards which is mentioned in a given sentence. A review of length $N$ can be represented as $X_r = \{w_1, w_2, \ldots, w_N\}$ where $w_i$ denotes $i^{th}$ word. To tackle this task, we consider it as a multi-label classification problem where the output can be defined as a binary vector $Y = \{y_1, y_2, \ldots, y_C\}$ where $y_c$ denotes $c^{th}$ aspect category, C is the number of aspect category of the specific domain.
- *Category-Sentiment Classification*: Given a review sentence of length $N$ can be represented as $X_r = \{w_1, w_2, \ldots, w_N\}$ where $w_i$ denotes $i^{th}$ word. The main task of CSC

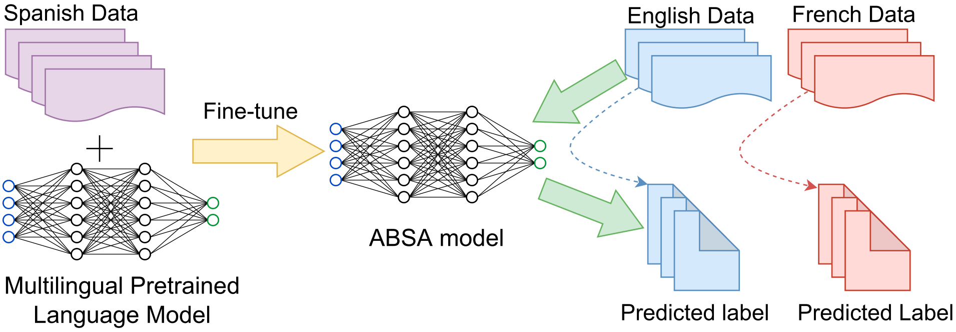

**Figure 1.** The architecture of zero-shot cross-lingual for ABSA tasks.

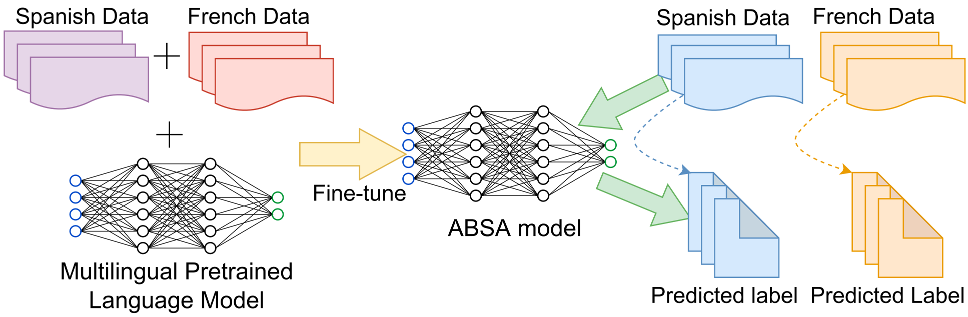

**Figure 2.** The architecture of joint training cross-lingual for ABSA tasks.

task is to detect the aspect categories and identify the associated sentiment polarities in the sentence. Formally, let the output Y be the set of one-hot vectors $Y = \{ap_1, ap_2, \ldots, ap_c)\}$ with $ap_i = (y_i^a, y_i^p)$ where $y_i^a$ represents the *i*-th category in the set of C aspect categories, $y_i^p$ represents the sentiment corresponding to the *i*-th aspect category in the set of positive, neutral, negative sentiment labels.

### 3.2. Methodology

In recent years, deep contextual language models have been introduced and the SOTA results have been achieved in various downstream NLP tasks. These models are already trained on a large unlabelled corpus and then is fine-tuned to downstream tasks. The aim of this study is to perform the zero-shot and joint training cross-lingual for ABSA tasks in five languages using transfer learning techniques based on pre-trained language models. However, we need a large amount of data and computational resources to train these models, which might not be possible for low-resource languages. Therefore, pre-trained multilingual language models are released to tackle this gap in research. From the work of Kalyan et al. (2021), there are many available multilingual language models. We employ the models mBERT and XLM-R because they support the languages in our experimental datasets. Moreover, we employ the latest SOTA multilingual models, such as InfoXLM and XLM-Align to conduct our experiments based on their ability in cross-lingual NLP tasks. Late in this section, we summarize the pre-trained multilingual transformer models used in this paper:

- *mBERT*: This is the BERT architecture Devlin et al. (2019) trained on a multilingual Wikipedia of 104 highest-resource languages on the two tasks: Masked language modelling (MLM) và Next sentence prediction (NSP).
- *XLM-R*: An optimized version of BERT which is trained based on the MLM task on 2.5T of data across 100 languages filtered from Common Crawl text (Conneau et al., 2020). This model outperforms the mBERT model in a variety of cross-lingual NLP tasks.
- *InfoXLM*: Chi et al. (2021a) presented a new cross-lingual pre-trained language model, named InfoXLM. This model is trained on monolingual and parallel data based on jointly training cross-lingual contrast with multilingual masked language modelling and translation language modelling. The pre-training data is similar to XLM-R model

(Conneau et al., 2020). The experimental results demonstrated that InfoXLM achieved better performance in cross-lingual transferability.

- *XLM-Align*: This is a pre-trained cross-lingual language model by applying the denoising word alignment task (Chi et al., 2021b). The model's training process consists of two steps: (1) self-labelling word alignments for translation pair; (2) random mask tokens in the bitext sentence. The extensive experiments on cross-lingual tasks showed that this model is effective for various datasets such as Sentence Classification (XNLI, PAWS-X), Question Answering (XQuAD,MLQA, and TyDiQA), etc.

To explore the performance of zero-shot and joint training cross-lingual approaches in different languages, we fine-tuned the above models based on the recommendation of the previous work (Devlin et al., 2019) and use different additional linear layers for each task. The detail of the two based models is described in the following section.

### 3.2.1. Aspect category detection

Aspect Category Detection is a multi-label classification where zero or more aspect categories can be detected from the sentence. Figure 3 illustrates the architecture for this task. Let the input sentence consists of a sequence of words: $X = \{w_1, w_2, \ldots, w_N\}$ where $w_i$ denotes $i$th word. After pre-processing text, two special tokens noted [CLS]

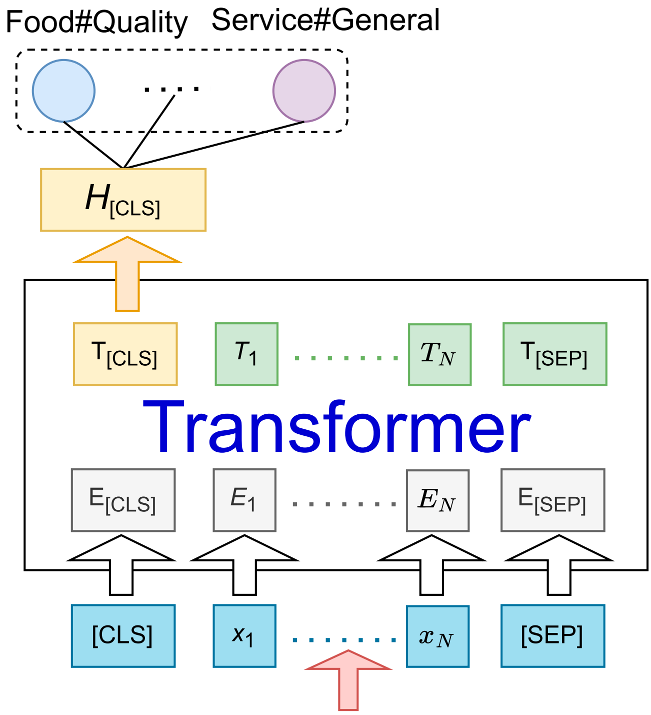

**Figure 3.** The overall architecture is based on the pre-trained transformer language models for the Aspect Category Detection task.

and [SEP] are added to the beginning and ending of the sequence. Because the experimental data is the sentence-level review, we use the padding operation to pad sentences in a uniform length. The max length value is the length of the longest sentence in the data set and is ensured to be shorter than the input of the transformer models. Then, this sequence is fed directly to pre-trained multilingual language models to obtain the representations for tokens. The output is a sequence of hidden states ($H_X^L$) represented as follows:

$$H_X^L = [h_{CLS}^L, h_1^L, h_2^L, \ldots, h_{SEP}^L] \tag{1}$$

where $H_X^L \in R^{N \times dim_h}$, $dim_h$ is the dimension of the representation vector, $h_i^L$ is the hidden state of $i^{th}$ input token in L transformer layers. The final hidden state $h_{CLS}^L$ of the [CLS] token in the last layer is used as the representation of input review. Finally, a fully connected layer with a sigmoid activation is added to the top model for task-specific. The model's output is a probability vector for the length corresponding to the size of the number of pre-defined categories. The sigmoid function will generate the corresponding probability of whole aspect categories. The aspect category is assigned to the review if the probability is greater than a threshold. The threshold is optimized on the validation set by using grid search. We employ the binary cross entropy as the loss function to calculate the predicted probability with the true label:

$$L(\Theta) = -\sum_{i=1}^{C} y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i) \tag{2}$$

### 3.2.2. Category-Sentiment classification

The output of this task is the pairs of {Aspect category, Sentiment} mentioned in a given review. It means that this compound task detects the aspect categories and their corresponding sentiment polarities simultaneously in this compound task. To deal with this problem, a multi-task approach based on the BERT models, inspired by the previous works (Dai et al., 2019; Schmitt et al., 2018; Van Thin et al., 2022), is employed to predict the output of each aspect category with its corresponding sentiment as a one-hot vector with four elements. The first element indicates whether the aspect category is mentioned in the review, while the three other elements represent three levels of sentiment polarity of each category; for example, the pair of 'Quality, positive' is encoded as [0 1 0 0]. As shown in Figure 4, we have the C softmax output layers corresponding to C aspect categories. We can train a model for an aspect category independently; however, this does not help the model explore correlated information between categories. Therefore, we build a multi-task architecture to utilize the correlation and influence between multi-aspect categories in the review. As similar to the ACD architecture, we use the last hidden state of the CLS token $H_{cls}^L$ as the representation of input review and feed it into the *C* fully connected layers with softmax activation.

$$\hat{y}^{(a)} = Softmax(W^{(a)} \cdot H_{cls}^L + b^{(a)}) \tag{3}$$

where *a* is the aspect category $a^{th}$ in the total C aspect categories, weight *W* and bias *b* are the parameters during training. Our model is optimized by minimizing the sum of

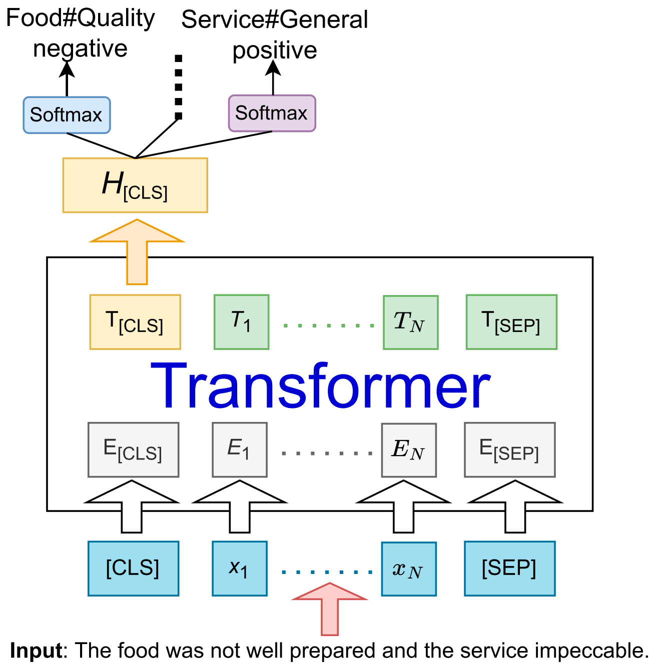

**Figure 4.** The overall architecture is based on the pre-trained transformer language models for the Category-Sentiment Classification task.

categorical cross-entropy loss in each category as follows:

$$L(\Theta) = -\sum_{a\in C}\sum_{i=1}^{4} y_i^a \cdot \log\left(\hat{y}_i^a\right) \tag{4}$$

where C is the number of the aspect category, $y_i^a$ is the true one-hot vector for the *a* category and $\hat{y}_i^a$ is the probability vector of prediction for the *a* category.

### 3.3. Zero-shot cross-lingual transfer learning

In most studies in the NLP field, termed zero-shot cross-lingual transfer learning means that the transfer model which is trained on the source language can be used to predict the target language without training data (Karthikeyan et al., 2019; Keung et al., 2019, 2020; Lauscher et al., 2020a; Nooralahzadeh et al., 2020). Based on this scenario, we experiment with two steps as follows: (1) Fine-tuning pre-trained multilingual language models on the training data of source language; (2) transferring the knowledge weight to evaluate the test data of the target language. Figure 5 shows this strategy for the zero-shot cross-lingual evaluation between two languages in our experiments. Unlike previous works, in this paper, we train and transfer the model to different source languages instead of using only English. Furthermore, we evaluated the performance when using the combination of multiple source languages as training data.

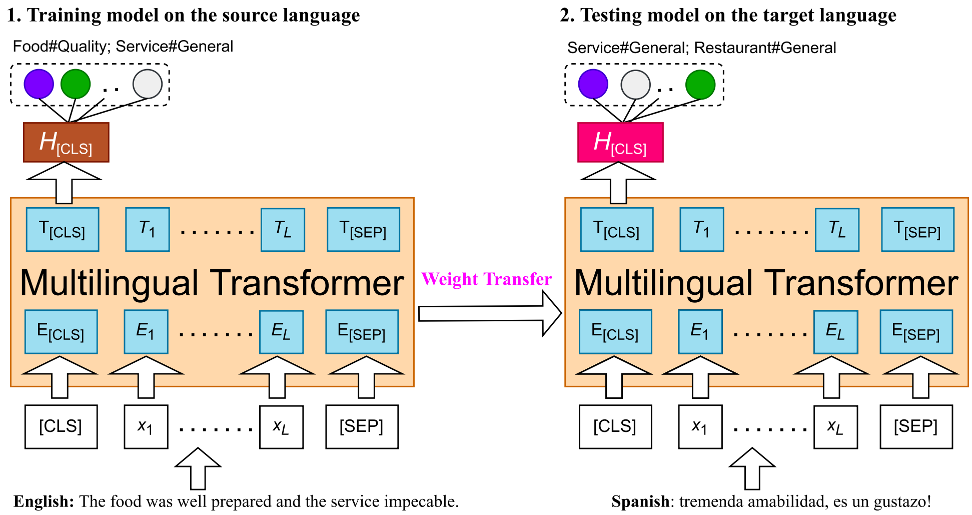
1. Training model on the source language
Food#Quality; Service#General
$H_{[CLS]}$
$T_{[CLS]}$ $T_1$ ....... $T_L$ $T_{[SEP]}$
Multilingual Transformer
$E_{[CLS]}$ $E_1$ ....... $E_L$ $E_{[SEP]}$
[CLS] $x_1$ ....... $x_L$ [SEP]
English: The food was well prepared and the service impecable.
Weight Transfer
2. Testing model on the target language
Service#General; Restaurant#General
$H_{[CLS]}$
$T_{[CLS]}$ $T_1$ ....... $T_L$ $T_{[SEP]}$
Multilingual Transformer
$E_{[CLS]}$ $E_1$ ....... $E_L$ $E_{[SEP]}$
[CLS] $x_1$ ....... $x_L$ [SEP]
Spanish: tremenda amabilidad, es un gustazo!

**Figure 5.** The weight transfer strategy between two languages for zero-shot cross-lingual evaluation for the Aspect Category Detection task.

### 3.4. Cross-lingual joint training

In zero-shot learning, the model is trained and tested on two different languages, while in multilingual joint learning, the model is trained on the combination of source and target language data. For example, we have labelled training data in English and French language as $L_{en}$ and $L_{fr}$. The task is to use $L_{en}$ and $L_{fr}$ to train a model and classify the review texts in the target language $L_{fr}$. This approach has been shown to be beneficial in various cross-lingual tasks (Aharoni et al., 2019; Johnson et al., 2017; Mulcaire et al., 2018; van der Heijden et al., 2020; Zhou et al., 2016). In order to evaluate the benefit of joint training based on multilingual language model, we conduct two experiments for the ABSA task as follows: (1) Combining the full training data of source and target language pair as the new training set and then evaluate on the test set of the target language, (2) Combining the entire training data of all languages to train a model and then evaluating it on the test sets of each language.

## 4. Experiments

### 4.1. Datasets and settings

#### 4.1.1. Datasets

We use the SemEval 2016 dataset (Pontiki et al., 2016) for the restaurant domain to conduct whole experiments, including languages such as English (en), French (fr), Spanish (es), Dutch (nl), and Russian (ru). These datasets have different sizes and this difference greatly affects the experimental results from the zero-shot and joint learning scenarios, especially in languages with lots of training data. To address this challenge, we use iterative stratification (Sechidis et al., 2011) to recreate the datasets for our experiments. These datasets are split into sub-datasets, including about 1200 training samples and 400 testing samples for five languages. The statistics of datasets are shown in Table 1.

**Table 1.** Distribution of datasets for five languages in our experiments.

| Aspect category | English | | French | | Dutch | | Spanish | | Russian | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Train | Test | Train | Test | Train | Test | Train | Test | Train | Test |
| Restaurant#General | 294 | 99 | 298 | 99 | 293 | 100 | 388 | 136 | 289 | 99 |
| Service#General | 295 | 99 | 344 | 115 | 373 | 127 | 347 | 122 | 406 | 138 |
| Food#Quality | 474 | 159 | 430 | 143 | 406 | 139 | 467 | 164 | 292 | 100 |
| Food#Style_Options | 92 | 31 | 183 | 61 | 126 | 43 | 121 | 43 | 99 | 34 |
| Drinks#Style_Options | 22 | 7 | 18 | 6 | 26 | 9 | 20 | 7 | 15 | 5 |
| Drinks#Prices | 12 | 4 | 13 | 4 | 14 | 4 | 11 | 4 | 5 | 2 |
| Restaurant#Prices | 53 | 18 | 71 | 24 | 48 | 17 | 80 | 28 | 32 | 11 |
| Restaurant#Miscellaneous | 68 | 23 | 79 | 26 | 17 | 6 | 14 | 5 | 18 | 6 |
| Ambience#General | 148 | 49 | 152 | 51 | 156 | 53 | 194 | 68 | 219 | 75 |
| Food#Prices | 54 | 18 | 62 | 21 | 37 | 12 | 85 | 30 | 18 | 6 |
| Location#General | 21 | 7 | 36 | 12 | 22 | 8 | 17 | 6 | 15 | 5 |
| Drink#Quality | 35 | 12 | 38 | 13 | 45 | 15 | 20 | 7 | 36 | 12 |

### *4.1.2. Experimental settings*

In this study, we utilize the pre-trained multilingual language models, which are available on the Hugging Face library (Wolf et al., 2020), including mBERT[1], XLM-R[2], XLM-Align[3], and InfoXLM[4] for the base version, XLM-R[5] and InfoXLMInfoXLM[6] for the large version. For the zero-shot and joint learning cross-lingual evaluation, we use the best individual-based model due to the limitation of computational resources. For the hyper-parameters, this study applies the cross-validation technique on the training set to choose the optimized parameters for each model and language. These parameters are learning rate, number of epochs, etc., mentioned in Table 2. For both tasks, we implemented an AdamW optimizer for different learning rates for each language. Batch sizes are selected with 32, 64, and 16. The experiments have shown that our model requires more epochs to prevent the underfitting problem because the two tasks are the type of the multi-label classification task. Therefore, we set the number of epochs at 30 and 40 for the ACD and CSC tasks, respectively, with an early stopping strategy.

### *4.1.3. Evaluation metrics*

The experimental results are reported using micro-averaging $F1-score_{Micro}$ by calculating the model's output to the gold annotation (Pontiki et al., 2016, 2015). The $F1-score_{Micro}$ is the common evaluation metric to evaluate the performance of ABSA task. To get the $F1-score_{Micro}$, we calculate the $Precision_{Micro}$ and $Recall_{Micro}$ for each

**Table 2.** Hyper-parameters settings of two tasks for the study.

| Hyperparameters | Value | |
|---|---|---|
| | ACD | CSC |
| Layers | 12 | 12 |
| Batch size | 32, 64 | 32 |
| Adam ε | 1e−6 | 1e−6 |
| Adam $\beta$ | (0.9,0.98) | (0.9,0.98) |
| Learning rate | 5e−5, 7e−5 | 2e−5, 5e−5 |
| Learning rate schedule | Linear | Linear |
| Warmup steps | 10,000 | 10,000 |
| Gradient clipping | 1.0 | 1.0 |
| Weight decay | 0.01 | 0.01 |
| Epochs | 30 | 40 |

class as follows:

$$Precision_{Micro} = \frac{\sum_{i=1}^{C} T_P^i}{\sum_{i=1}^{C} (T_P^i + F_N^i)} \tag{5}$$

$$Recall_{Micro} = \frac{\sum_{i=1}^{C} T_P^i}{\sum_{i=1}^{C} (T_P^i + F_P^i)} \tag{6}$$

$$F1 - score_{Micro} = \frac{2 \times Precision_{Micro} \times Recall_{Micro}}{Precision_{Micro} + Recall_{Micro}} \tag{7}$$

where $T_P$ represents the number of instances that are correctly predicted with positive label, while $F_P$ and $F_N$ represent the number of instances with labels incorrectly predicted to be positive or negative, respectively. The value $C$ shows the number of classes; in our case $C = 12$. For the ACD task, the classes are the Entity-Attribute pairs, while the F1-score will be calculated based on the tuples (Entity-Attribute-Polarity) for the CSC task.

### 4.2. Results

This section presents the experimental results of three scenarios: (1) performance of different pre-trained language models; (2) evaluation of zero-shot learning; (3) evaluation of joint learning scenario.

#### 4.2.1. Performance of multilingual models

We first compare the effectiveness of multilingual pre-trained language models on two ABSA tasks. Tables 3 and 4 show the $F1 - score_{Micro}$ per model per language for the tasks of ACD and CSC, with the highest scores per language shown in bold and underlined. Note that the results represent the performance of models trained and tested on the specific language data. It can be observed that the large models improve the results over base models further. $InfoXLM_{large}$ model achieved the highest scores in most languages except Dutch for the ACD task. While $InfoXLM_{large}$ model shows the effectiveness in all languages for the CSC task. Our experimental results confirmed that the large models perform better than base pre-trained models. Among based language models, the XLM-Align model achieved high scores for languages except Dutch and Russian for the ACD task. One of the reasons for the worse performance of the XLM-Align model in comparison with the XLM-R model in Dutch and Rusian is the size of pre-training data in the pre-trained model as a previous study (Lauscher et al., 2020b). Specifically, the size of pre-training data of the XLM-R model is larger than XLM-Align

**Table 3.** The performances of multilingual pre-trained language models on five datasets for Aspect Category Detection task.

| | Architecture | English | French | Dutch | Spanish | Russian |
|---|---|---|---|---|---|---|
| Base model | mBERT | 73.33 | 67.57 | 72.80 | 76.86 | 81.01 |
| | XLM-R | 81.23 | 71.76 | **80.38** | 79.05 | **83.81** |
| | InfoXLM | 81.20 | 76.51 | 79.43 | 79.30 | 83.50 |
| | XLM-Align | **81.62** | **77.57** | 77.05 | **81.03** | 83.61 |
| Large model | XML-R | 82.99 | 80.14 | **83.77** | 82.37 | 87.50 |
| | InfoXLM | **83.69** | **80.31** | 83.57 | **83.01** | **87.80** |

**Table 4.** The performances of multilingual pre-trained language models on five datasets for Category Sentiment Classification task.

| | Architecture | English | French | Dutch | Spanish | Russian |
|---|---|---|---|---|---|---|
| Base model | mBERT | 52.86 | 53.75 | 57.84 | 61.56 | 58.44 |
| | XLM-R | 67.06 | 54.89 | **63.79** | 68.66 | **68.26** |
| | InfoXLM | 68.24 | 60.14 | 63.37 | 68.03 | 67.77 |
| | XLM-Align | **69.76** | **61.86** | 63.59 | **68.65** | 67.58 |
| Large model | XML-R | 69.44 | 67.77 | 67.77 | 72.58 | 73.01 |
| | InfoXLM | **70.42** | **69.16** | **70.33** | **73.74** | **73.80** |

(29.3G and 25.9G for Dutch, 278G, and 253.3G for Russian). Compared with $XLM - R_{base}$, XLM-Align gives 0.4%, +5.81%, and 1.98% improvements in English, French, and Spanish. While XLM-R still shows effectiveness in the Dutch and Russian with 3.33% and 0.2% improvements than XLM-Align. Comparing the results of InfoXLM against XLM-Align shows that the performance of the two models is quite competitive, which depends on the language and task, but the difference is not significant. These experimental results are similar to the CSC task. These results show that the performance of pre-trained multilingual language models is different based on the language. Therefore, we recommend that future studies compare the performance of these models in a specific language, particularly low-resource languages, to select the best model.

### *4.2.2. Zero-shot cross-lingual*

In this section, we present the results in zero-shot on two following scenarios. First, we investigate the performance of each model, which is trained on the source language and evaluated on the target language. Secondly, we examine the effectiveness of the model which are trained on multiple source languages without the target language. We choose the XLM-Align as the primary model for this setting.

Table 5 presents the zero-shot cross-lingual results on source-target language pairs of the ACD and CSC tasks. The row and column represent the source and target language, respectively. The best performance of the cross-lingual language pair is bold for each column. As shown in Table 5, it can be seen that training model on the English language achieves the best scores on the French language for two tasks. Moreover, we observe that the performances of the XLM-Align model are different on two tasks in some language pairs in this setting. For example, using French as the source language gives the best score on Dutch for the ACD task; however, the best source language for the CSC task is English. It can be seen that the CSC task is a compound task that aim to assign the set of aspect categories and corresponding sentiment – which may yield different results than ACD. In addition, the ratio of training data toward the sentiment polarity class

**Table 5.** The results of zero-shot cross lingual on the five datasets using XLM-Align model.

| Language | Aspect category detection | | | | | Category sentiment classification | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | en | fr | nl | es | ru | en | fr | nl | es | ru |
| en | – | **71.08** | **71.70** | 72.84 | 76.69 | – | **53.74** | **57.52** | **61.88** | 60.00 |
| fr | 74.02 | – | 70.57 | **73.11** | 76.53 | 58.41 | – | 56.86 | 59.85 | 57.52 |
| du | **74.50** | 69.72 | – | 70.11 | 76.14 | **62.81** | 53.66 | – | 61.32 | 60.47 |
| nl | 70.23 | 69.86 | 70.13 | – | **79.92** | 60.10 | 53.35 | 56.50 | – | **61.25** |
| ru | 70.76 | 68.70 | 70.30 | 72.09 | – | 56.35 | 44.74 | 51.35 | 55.56 | – |

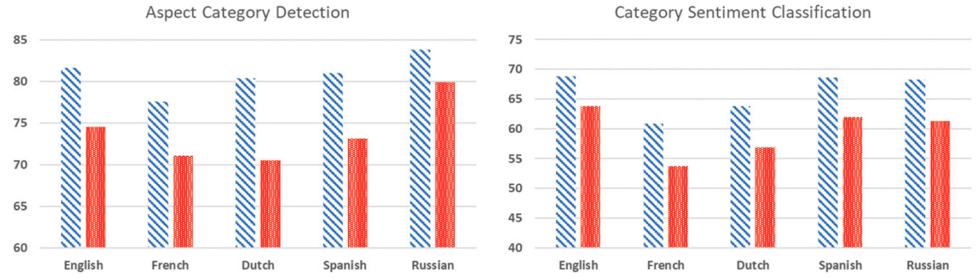

**Figure 6.** The graph compares the best results in monolingual models, which are trained both trained and tested on target language data (blue bar), and zero-shot cross-lingual setting (red bar).

between languages is also different. This leads to inconsistent experimental results between the two tasks. Moreover, our experimental results indicate the effect of similar languages on zero-shot transfer learning. For example, the model is trained on English (a Germanic language) data to produce the higher scores in typologically or etymologically languages such as French and Spanish (Romance languages) than Russian (a Slavic language). The reason is that most of modern English vocabulary is borrowed from the Romance languages (Şenel et al., 2017).

This demonstrates that the selection of source language to transfer knowledge in the zero-shot setting also influences performance in the target language. Figure 6 shows the best scores of the monolingual model compared with the zero-shot cross-lingual setting. In general, it is obvious that the performance of zero-shot learning is lower than the monolingual model. This result is acceptable in the context of no training data in the target language. Specifically, when comparing the highest zero-shot cross-lingual results with the monolingual results, we can see that the difference ranges from 3.89% to 9.81% and 4.96% to 7.12% for the ACD and CSC tasks, respectively.

Moreover, we conduct an experiment to find out the effectiveness of the training model in a combination of multiple languages. Therefore, in this experiment, we combine the training data of source languages without the target language. Table 6 presents the scores of this experiment for two tasks. We can also observe that the training model on combination data gains better scores in all of the languages in both tasks. In particular, there is a remarkable increase for the ACD task in Dutch, Spanish, and English with +7.33%, +5.25%, and +4.41%, respectively. For the CSC task, the model also improves $F1 - score_{Micro}$ in the range from 3.35% to 7.67% in all languages.

**Table 6.** The results of zero-shot setting where training XLM-Align model on a combination of different languages.

| Source ⇒ Target | ACD Task | | | CSC Task | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | F1-score | Precision | Recall | F1-score |
| fr + ru + nl + es ⇒ en | 81.12 | 76.81 | 78.91 | 71.14 | 67.49 | 69.27 |
| en + nl + es + ru ⇒ fr | 75.56 | 69.91 | 72.63 | 64.08 | 58.96 | 61.41 |
| en + fr + es + ru ⇒ nl | 77.76 | 78.05 | 77.90 | 63.62 | 63.98 | 62.81 |
| en + fr + nl + ru ⇒ es | 75.49 | 81.45 | 78.36 | 66.09 | 68.23 | 67.14 |
| en + fr + nl + es ⇒ ru | 78.88 | 82.56 | 80.67 | 65.96 | 63.29 | 64.60 |

**Table 7.** The results of joint training of source and target languages on the five dataset.

| Language | Aspect category detection | | | | | Category sentiment classification | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | en | fr | nl | es | ru | en | fr | nl | es | ru |
| en | 81.62 | 77.33 | **80.70** | 81.62 | **86.15** | 69.76 | 64.74 | **68.51** | 68.68 | 71.38 |
| fr | 81.03 | 77.57 | 80.66 | **81.86** | 85.80 | 71.23 | 61.86 | 67.92 | **69.98** | 70.29 |
| nl | **82.61** | 77.17 | 80.38 | 81.49 | 85.60 | **73.35** | 63.59 | 63.79 | 69.02 | 69.33 |
| es | 82.32 | **78.78** | 80.52 | 81.03 | 86.03 | 72.32 | **65.50** | 67.80 | 68.65 | **72.20** |
| ru | 80.43 | 75.76 | 78.84 | 81.08 | 83.81 | 73.03 | 60.78 | 65.27 | 69.84 | 68.26 |

Note: * Bold value represents the best score of the target language.

### 4.2.3. Joint training

Table 7 shows the results of joint training of language pairs for two tasks. The main diagonal represents the best scores (Tables 3 and 4) where the model is trained and tested on the data of the target language. In general, the joint training approach can improve the performance of the model; however, it is obvious that the improvements depend on the language pair. We found that training the pairs in the same language group related to linguistic relations[7] such as English with Dutch, French, and Spanish brings the benefit over other pairs. In addition, instead of joint training for each language pair, we combine the training data for all languages, including the target language, as the final training set. The results of this experiment are summarized in Table 8.

## 4.3. Discussion

First, we discuss the role of source languages in zero-shot cross-lingual transfer learning based on our experimental results. We surveyed that most previous studies (Larochelle et al., 2008) choose English as the source language to transfer knowledge in zero-shot learning because of the following reasons: (1) English is one of the rich-resource languages in research community (Joshi et al., 2020b); (2) English is a top language that have a large size in the pre-training data in most current pre-trained multilingual language models (Chi et al., 2021a, 2021b; Conneau et al., 2020; Devlin et al., 2019). However, our experimental results, which are shown in Table 5 indicated that other

**Table 8.** The results of joint training approach based on the combination of multiple training sets.

| Sources ⇒ target | ACD task | | | CSD task | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | F1-score | Precision | Recall | F1-score |
| en + fr + nl + es + ru ⇒ en | 83.49 | 83.65 | 83.57 | 76.13 | 73.95 | 75.02 |
| en + fr + nl + es + ru ⇒ fr | 78.31 | 79.13 | 78.72 | 68.50 | 65.04 | 66.73 |
| en + fr + nl + es + ru ⇒ nl | 78.78 | 82.18 | 80.44 | 70.70 | 72.42 | 71.55 |
| en + fr + nl + es + ru ⇒ es | 80.00 | 84.52 | 82.20 | 70.63 | 71.77 | 71.20 |
| en + fr + nl + es + ru ⇒ ru | 87.25 | 87.42 | 87.34 | 74.22 | 72.41 | 73.31 |

**Table 9.** Results of three approaches on five languages for the aspect category detection.

| Approach | | en | fr | nl | es | ru |
|---|---|---|---|---|---|---|
| Monolingual learning | target ⇒ target | 81.62 | 77.57 | 80.38 | 81.03 | 83.81 |
| Zero-shot learning | source ⇒ target | 74.02 | 71.08 | 70.57 | 73.11 | 79.92 |
| | multiple sources ⇒ target | 78.91 | 72.63 | 77.90 | 78.36 | 80.67 |
| Joint training | source + target ⇒ target | 82.62 | 78.78 | 80.70 | 81.86 | 86.15 |
| | multiple sources + target ⇒ target | 83.57 | 78.72 | 80.44 | 82.20 | 87.34 |

**Table 10.** Results of three approaches on five languages for the Category Sentiment Classification task.

| Approach | | en | fr | nl | es | ru |
|---|---|---|---|---|---|---|
| Monolingual learning | target ⇒ target | 69.76 | 61.86 | 63.79 | 68.65 | 68.26 |
| Zero-shot learning | source ⇒ target | 62.81 | 53.74 | 56.86 | 61.88 | 61.25 |
| | multiple sources ⇒ target | 69.27 | 61.41 | 62.81 | 67.14 | 64.60 |
| Joint training | source + target ⇒ target | 73.35 | 65.50 | 68.51 | 69.98 | 72.20 |
| | multiple sources + target ⇒ target | 75.02 | 66.73 | 71.55 | 71.20 | 73.31 |

languages could be the source language instead of English. For example, transferring knowledge from French produced the highest scores in Spanish for the ACD task, while Spanish gave the best score for the Russian language for both tasks. In addition, it is observed that the languages which belong to the same language group with closer linguistic relations, such as English and Dutch, Spanish and French, or with close lexical distance (Şenel et al., 2017), such as English and French, produced the highest scores for the target language in the zero-shot setting. Therefore, the linguistic relationship and lexical distance between the source language and the target language play an important role in zero-shot cross-lingual learning.

Second, Tables 9 and 10 show the performances of three approaches for the ACD and CSC tasks, respectively. Generally, we observe that the results of zero-shot learning – which is trained on the source language and tested on the target language, are always lower than the results of monolingual learning – which is trained and tested on the target language. Another interesting point is that training model on multiple source languages improves the performance in terms of F1-score than a source language for all languages in both tasks. Especially, we notice that the performance of zero-shot learning in multiple source settings gives approximately the best results than the monolingual setting for the CSC task in all languages except Russian. The reason might relate to the distance between languages and the distribution of aspect categories with corresponding sentiment polarity. Because the CSC task considers the category and polarity as ground truth and has a large imbalance between the classes, therefore, combing multiple source languages can increase and diversify the number of training samples. When comparing the joint training results with the remaining, it can be seen that joint training improves the performance of the model in all languages, particularly close-relation languages. Furthermore, by combining multiple training data for all languages, including the target language, the model performs better than using only one source language for all languages; the difference is significant for the CSC task.

In order to explore the effectiveness of the joint learning approach, we conducted experiments with different sizes of the training set in the source and target language. We consider two scenarios: (1) combining the full training set of the source language with part of the training set of the target language; (2) In contrast, we combine a part of the training set of the source language with the full training set of the target language. The source language for a specific language is selected based on joint learning results (see in Table 7). For example, the combination of English and Dutch languages produces the best score for the Dutch language. As shown in Figures 7 and 8, we can see that the model's performance increases with the size of the training samples for the target language. These results proved that combining more annotated data for the target

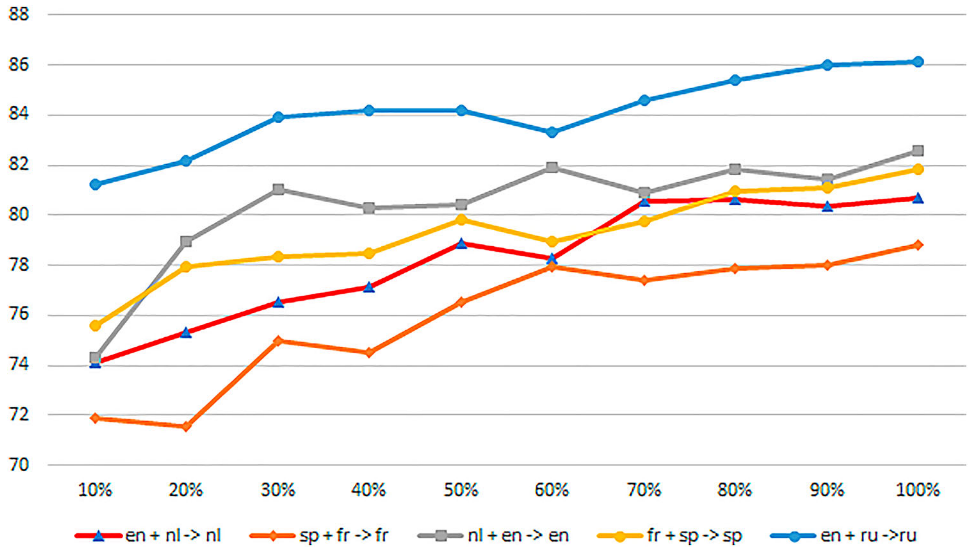

**Figure 7.** Joint learning results for the combination of the source language with the amount of training samples from the target language for the Aspect Category Detection task.

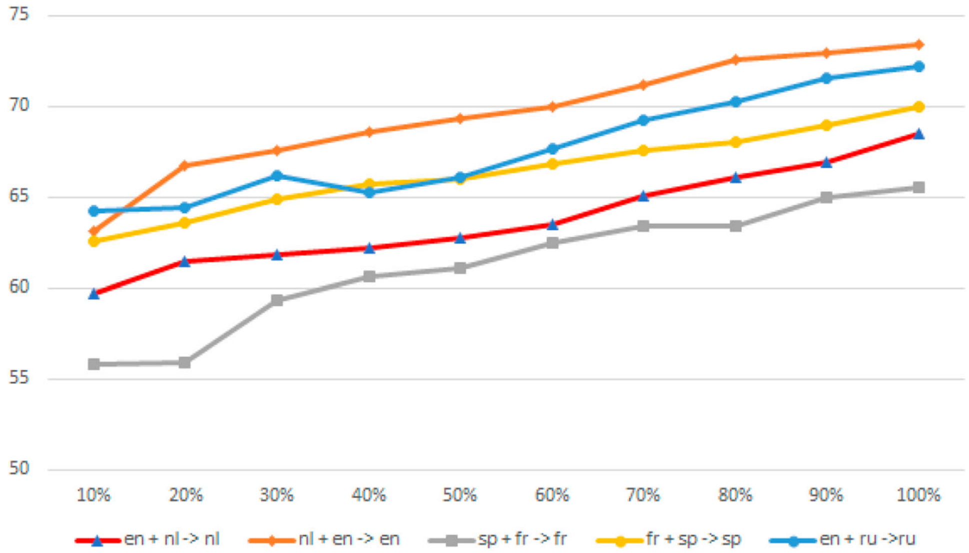

**Figure 8.** Joint learning results for the combination of the source language with amount of training samples from the target language for the Category-Sentiment Classification task.

language increases the performance of multilingual models, which are only trained on the source language data.

## 5. Conclusion

In this paper, we studied the ability of different contextualized multilingual language models in the zero-shot and joint training cross-lingual settings. We conducted experiments on two sub-tasks in the ABSA problem for five languages. For the zero-shot cross-lingual setting, we explore two scenarios relying on two strategies: (1) training on a source language; (2) training on the multiple source languages to fine-tuning the models. The results showed that it is beneficial to take advantage of multiple source languages in a zero-shot cross-lingual setting. Moreover, the experimental results indicate that the selection of source languages also plays an important role in achieve good results for the target language. Although our results indicated that the performance of zero-shot learning is not as good as in a monolingual setting, the results are pretty impressive in case there is no training data for the target language.

For the joint training cross-lingual study, two experiments were conducted to demonstrate the effectiveness of the multilingual language models on the mixture dataset. We found that a joint training model on the group languages with linguistic relations can perform better than monolingual data. Through extensive experiments in several languages, we demonstrated the efficacy of cross-lingual joint training. Furthermore, we explored the combination of the source language with amount of training samples from the target language. The results indicated that the performance of the model increased proportionally to the number of data samples in the target language.

Finally, from the performance of fine-tuning various pre-trained multilingual language models in five languages, we recommend that future studies should compare the performance of different models in the specific language in order to select the best model, especially for low-resource languages. In future work, we plan to examine a broader set of languages (Asian, Africa, etc.) and tasks to obtain more comprehensive evaluations.

## Notes

1. https://huggingface.co/bert-base-multilingual-cased
2. https://huggingface.co/xlm-roberta-base
3. https://huggingface.co/microsoft/xlm-align-base
4. https://huggingface.co/microsoft/infoxlm-base
5. https://huggingface.co/xlm-roberta-large
6. https://huggingface.co/microsoft/infoxlm-large
7. https://termcoord.eu/2014/01/lexical-distance-languages-europe/

## Acknowledgments

Dang Van Thin proposed the main ideas of this research, conducted the experiments and wrote the paper. Hung Q. Ngo, Duong Ngoc Hao and Ngan Luu-Thuy Nguyen discuss the ideas and contributed to writing the paper. All authors approved the final version.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

This research is funded by University of Information Technology-Vietnam National University HoChi-Minh City [grant number D1-2023-01].

## Notes on contributors

*Dang Van Thin* is currently PhD student at University of Information Technology - VNUHCM. He graduated with the Bachelor and Master degree in computer science at the University of Information Technology - Vietnam National University Ho Chi Minh city, Vietnam in 2017 and 2020, respectively. He also a member of Multimedia Communications Laboratory (MMLab) and his research interests are about natural language processing, machine learning, deep learning and applications.

*Hung Quoc Ngo* was awarded PhD degree in Computer Science from University College Dublin, Ireland. He received a Master degree in Computer Science from University of Science-VNUHCM, Vietnam. He is currently a lecturer with the School of Business Technology, Retail, and Supply Chain, Technological University Dublin, Ireland. He has involved in the BioCaster project by building geographical ontology, integrating the geo-ontology into the Global Health Monitor system, and building the webpage for publishing project results. Recently, he has built knowledge graphs for digital agriculture in CONSUS project at University College Dublin. His research interests are natural language processing, knowledge management, and data analytics.

*Dr Duong Ngoc Hao* is a lecturer of the department of Department of Maths and Physics at University of Information Technology - VNUHCM. He received the B.S. degree in Math - Informatics from HCMC University of Education and the M.S. degree in University of Science, Vietnam National University - Ho Chi Minh City, Vietnam. He got Ph.D. degree from Institute of Mechanics, Vietnam. His interests is math for Computer Science, Machine Learning Algorithm, and Natural Language Processing.

*Ngan Luu-Thuy Nguyen* is a scientist at the University of Information Technology, Vietnam National University, Ho Chi Minh City, Vietnam. She received her PhD degree in information science and technology from the University of Tokyo, Japan. She was a postdoctoral researcher at the National Institute of Informatics, Japan from 2012 to 2013. Her research interests include natural language processing and data analysis.

## ORCID

*Dang Van Thin* http://orcid.org/0000-0001-8340-1405

## References

Aharoni, R., Johnson, M., & Firat, O. (2019). Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 3874–3884). Minneapolis, Minnesota: Association for Computational Linguistics.

Ammar, W., Mulcaire, G., Ballesteros, M., Dyer, C., & Smith, N. A. (2016). Many languages, one parser. *Transactions of ACL*, *4*, 431–444. https://doi.org/10.1162/tacl_a_00109

Arkhipov, M., Trofimova, M., Kuratov, Y., & Sorokin, A. (2019). Tuning multilingual transformers for language-specific named entity recognition. In *Proceedings of BSNLP* (pp. 89–93). Association for Computational Linguistics.

Artetxe, M., & Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of ACL*, *7*, 597–610. https://doi.org/10.1162/tacl_a_00288

Bender, E. M. (2011). On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology*, *6*. https://doi.org/10.33011/lilt.v6i.1239

Bhatnagar, V., Kumar, P., & Bhattacharyya, P. (2022). Investigating hostile post detection in Hindi. *Neurocomputing*, *474*, 60–81. https://doi.org/10.1016/j.neucom.2021.11.096

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of ACL*, *5*, 135–146. https://doi.org/10.1162/tacl_a_00051

Chi, Z., Dong, L., Wei, F., Yang, N., Singhal, S., Wang, W., Song, X., Mao, X.-L., Huang, H., & Zhou, M. (2021a). InfoXLM: an information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of NAACL* (pp. 3576–3588). Association for Computational Linguistics.

Chi, Z., Dong, L., Zheng, B., Huang, S., Mao, X.-L., Huang, H., & Wei, F. (2021b). Improving pretrained cross-lingual language models via self-labeled word alignment. In *Proceedings of ACL and the 11th international joint conference on natural language processing* (pp. 3418–3430). Association for Computational Linguistics.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL* (pp. 8440–8451). Association for Computational Linguistics.

Conneau, A., & Lample, G. (2019). Cross-lingual language model pretraining. In *Proceedings of NeurIPS* (Vol. 32). Curran Associates Inc.

Dai, Z., Dai, W., Liu, Z., Rao, F., Chen, H., Zhang, G., Ding, Y., & Liu, J. (2019). Multi-task multi-head attention memory network for fine-grained sentiment analysis. In *Proceedings of the CCF NLPCC* (pp. 609–620). Springer, Cham.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL* (pp. 4171–4186). Association for Computational Linguistics.

Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., & Klakow, D. (2021). A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of NAACL* (pp. 2545–2568). Association for Computational Linguistics.

Hoang, M., Bihorac, O. A., & Rouces, J. (2019). Aspect-based sentiment analysis using bert. In *NEAL Proceedings of NoDaLiDa* (Vol. 167, pp. 187–196). Linköping University Electronic Press.

Jebbara, S., & Cimiano, P. (2019). Zero-shot cross-lingual opinion target extraction. In *Proceedings of NAACL* (pp. 2486–2495). Association for Computational Linguistics.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., & Dean, J. (2017). Google's multilingual neural machine translation system: enabling zero-shot translation. *Transactions of ACL*, *5*, 339–351. https://doi.org/10.1162/tacl_a_00065

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020a). The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 6282–6293). Association for Computational Linguistics.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020b). The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of ACL* (pp. 6282–6293). Association for Computational Linguistics.

Kalyan, K. S., Rajasekharan, A., & Sangeetha, S. (2021). Ammus: a survey of transformer-based pretrained models in natural language processing. arXiv preprint arXiv:2108.05542.

Karimi, A., Rossi, L., & Prati, A. (2021). Adversarial training for aspect-based sentiment analysis with bert. In *Proceedings of ICPR* (pp. 8797–8803). IEEE.

Karthikeyan, K., Wang, Z., Mayhew, S., & Roth, D. (2019). Cross-lingual ability of multilingual bert: an empirical study. In *ICLR*.

Keung, P., Lu, Y., & Bhardwaj, V. (2019). Adversarial learning with contextual embeddings for zero-resource cross-lingual classification and ner. In *Proceedings of EMNLP-IJNLP* (pp. 1355–1360). Association for Computational Linguistics.

Keung, P., Lu, Y., Salazar, J., & Bhardwaj, V. (2020). Don't use English dev: on the zero-shot cross-lingual evaluation of contextual embeddings. In *Proceedings of EMNLP* (pp. 549–554). Association for Computational Linguistics.

Kim, J., Choi, N., Lim, S., Kim, J., Chung, S., Woo, H., Song, M., & Choi, J. D. (2021). Analysis of zero-shot crosslingual learning between English and Korean for named entity recognition. In *Proceedings of MRL* (pp. 224–237). IEEE.

Kumar, A., & Albuquerque, V. H. C. (2021). Sentiment analysis using XLM-R transformer and zero-shot transfer learning on resource-poor indian language. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *20*(5), 1–13. https://doi.org/10.1145/3461764

Larochelle, H., Erhan, D., & Bengio, Y. (2008). Zero-data learning of new tasks. In *Proceedings of AAAI* (pp. 646–651). AAAI Press.

Lauscher, A., Ravishankar, V., Vulić, I., & Glavaš, G. (2020a). From zero to hero: on the limitations of zero-shot language transfer with multilingual transformers. In *Proceedings of EMNLP* (pp. 4483–4499). Association for Computational Linguistics.

Lauscher, A., Ravishankar, V., Vulić, I., & Glavaš, G. (2020b). From zero to hero: on the limitations of zero-shot language transfer with multilingual transformers. In *Proceedings of EMNLP* (pp. 4483–4499). Association for Computational Linguistics.

Li, Z., Wei, Y., Zhang, Y., Zhang, X., & Li, X. (2019). Exploiting coarse-to-fine task transfer for aspect-level sentiment classification. In *Proceedings of the AAAI* (Vol. 33, pp. 4253–4260). AAAI Press.

Lin, Y.-H., Chen, C.-Y., Lee, J., Li, Z., Zhang, Y., Xia, M., Rijhwani, S., He, J., Zhang, Z., Ma, X., Anastasopoulos, A., Littell, P., & Neubig, G. (2019). Choosing transfer languages for cross-lingual learning. In *Proceedings of ACL* (pp. 3125–3135). Association for Computational Linguistics.

Mulcaire, P., Swayamdipta, S., & Smith, N. A. (2018). Polyglot semantic role labeling. In *Proceedings of ACL* (pp. 667–672).

Nooralahzadeh, F., Bekoulis, G., Bjerva, J., & Augenstein, I. (2020). Zero-shot cross-lingual transfer with meta learning. In *Proceedings of EMNLP* (pp. 4547–4562). Association for Computational Linguistics.

Pamungkas, E. W., Basile, V., & Patti, V. (2021). A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection. *Information Processing & Management*, *58*(4), 102544. https://doi.org/10.1016/j.ipm.2021.102544

Pei, Y., Chen, S., Ke, Z., Silamu, W., & Guo, Q. (2022). AB-LaBSE: Uyghur sentiment analysis via the pre-training model with BiLSTM. *Applied Sciences*, *12*(3), 1182. https://doi.org/10.3390/app12031182

Phan, K. T. -K., Ngoc Hao, D., Thin, D. V., & Luu-Thuy Nguyen, N. (2021). Exploring zero-shot cross-lingual aspect-based sentiment analysis using pre-trained multilingual language models. In *Proceeding of MAPR* (pp. 1–6). IEEE.

Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual BERT? In *Proceedings of ACL* (pp. 4996–5001). Association for Computational Linguistics.

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., & Hoste, V. (2016). Semeval-2016 task 5: aspect based sentiment analysis. In *Proceedings of SemEval* (pp. 19–30). Association for Computational Linguistics.

Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., & Androutsopoulos, I. (2015). SemEval-2015 task 12: aspect based sentiment analysis. In *Proceedings of SemEval* (pp. 486–495). Association for Computational Linguistics.

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., & Manandhar, S. (2014). Semeval-2014 task 4: aspect based sentiment analysis. In *Proceedings of SemEval* (pp. 27–35). Association for Computational Linguistics.

Rietzler, A., Stabinger, S., Opitz, P., & Engl, S. (2020). Adapt or get left behind: domain adaptation through bert language model finetuning for aspect-target sentiment classification. In *Proceedings of LREC* (pp. 4933–4941). European Language Resources Association.

Sarkar, A., Reddy, S., & Iyengar, R. S. (2019). Zero-shot multilingual sentiment analysis using hierarchical attentive network and bert. In *Proceedings of NLPIR* (pp. 49–56). Association for Computing Machinery.

Schmitt, M., Steinheber, S., Schreiber, K., & Roth, B. (2018). Joint aspect and polarity classification for aspect-based sentiment analysis with end-to-end neural networks. In *Proceedings of EMNLP* (pp. 1109–1114). Association for Computational Linguistics.

Sechidis, K., Tsoumakas, G., & Vlahavas, I. (2011). On the stratification of multi-label data. In *Proceedings of ECML-PKDD* (pp. 145–158). Springer.

Şenel, L. K., Yücesoy, V., Koç, A., & Çukur, T. (2017). Measuring cross-lingual semantic similarity across European languages. In *Proceedings of TSP* (pp. 359–363). IEEE.

Sharma, A., Kabra, A., & Jain, M. (2022a). Ceasing hate with MoH: hate speech detection in Hindi–English code-switched language. *Information Processing & Management*, *59*(1), 102760. https://doi.org/10.1016/j.ipm.2021.102760

Sharma, R., Morwal, S., & Agarwal, B. (2022b). Named entity recognition using neural Language model and CRF for Hindi Language. *Computer Speech & Language*, *74*, 101356. https://doi.org/10.1016/j.csl.2022.101356

Song, Y., Wang, J., Liang, Z., Liu, Z., & Jiang, T. (2020). Utilizing bert intermediate layers for aspect based sentiment analysis and natural language inference. arXiv preprint arXiv:2002.04815.

Sultan, A., Salim, M., Gaber, A., & El Hosary, I. (2020). WESSA at SemEval-2020 task 9: code-mixed sentiment analysis using transformers. In *Proceedings of SemEval* (pp. 1342–1347). International Committee for Computational Linguistics.

Sun, C., Huang, L., & Qiu, X. (2019). Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence. In *Proceedings of NAACL* (pp. 380–385). Association for Computational Linguistics.

Sun, X., Ge, T., Ma, S., Li, J., Wei, F., & Wang, H. (2022). A unified strategy for multilingual grammatical error correction with pre-trained cross-lingual language model. In *Proceedings of IJCNN* (pp. 4367–4374). Association for Computational Linguistics.

van der Heijden, N., Abnar, S., & Shutova, E. (2020). A comparison of architectures and pretraining methods for contextualized multilingual word embeddings. In *Proceedings of AAAI* (Vol. 34, pp. 9090–9097). AAAI Press.

Van Thin, D., Le, L. S., Nguyen, H. M., & Nguyen, N. L.-T. (2022). A joint multi-task architecture for document-level aspect-based sentiment analysis in vietnamese. *IJMLC*, *12*(4), 126–136. https://doi.org/10.18178/ijmlc.2022.12.4.1091

Wan, H., Yang, Y., Du, J., Liu, Y., Qi, K., & Pan, J. Z. (2020). Target-aspect-sentiment joint detection for aspect-based sentiment analysis. In *Proceedings of the AAAI* (Vol. 34, pp. 9122–9129). AAAI Press.

Wang, W., Zheng, V. W., Yu, H., & Miao, C. (2019). A survey of zero-shot learning: settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology*, *10*(2), 1–37. https://doi.org/10.1145/3293318

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., & Davison, J. (2020). Transformers: state-of-the-art natural language processing. In *Proceedings of EMNLP* (pp. 38–45). Association for Computational Linguistics.

Xu, H., Liu, B., Shu, L., & Philip, S. Y. (2019). Bert post-training for review reading comprehension and aspect-based sentiment analysis. In *Proceedings of NAACL* (pp. 2324–2335). Association for Computational Linguistics.

Zhou, X., Wan, X., & Xiao, J. (2016). Attention-based lstm network for cross-lingual sentiment classification. In *Proceedings of EMNLP* (pp. 247–256). Association for Computational Linguistics.