English-Malay Word Embeddings Alignment for Cross-lingual Emotion Classification with Hierarchical Attention Network

Lim Ying Hao, Jasy Liew Suet Yan
School of Computer Sciences, Universiti Sains Malaysia
11800 Penang, Malaysia
yinghaoly@student.usm.my, jasyliew@usm.my

Abstract

The main challenge in English-Malay cross-lingual emotion classification is that there are no Malay training emotion corpora. Given that machine translation could fall short in contextually complex tweets, we only limited machine translation to the word level. In this paper, we bridge the language gap between English and Malay through cross-lingual word embeddings constructed using singular value decomposition. We pre-trained our hierarchical attention model using English tweets and fine-tuned it using a set of gold standard Malay tweets. Our model uses significantly less computational resources compared to the language models. Experimental results show that the performance of our model is better than mBERT in zero-shot learning by 2.4% and Malay BERT by 0.8% when a limited number of Malay tweets is available. In exchange for 6 – 7 times less in computational time, our model only lags behind mBERT and XLM-RoBERTa by a margin of 0.9 – 4.3 % in few-shot learning. Also, the word-level attention could be transferred to the Malay tweets accurately using the cross-lingual word embeddings.

1 Introduction

Sentiment analysis and opinion mining are used interchangeably to represent the task of classifying the sentiment polarity of opinionated text (Meng et al., 2012). On a coarse-grained level, the task is often a binary classification problem (positive or negative) (Pang & Lee, 2005). The neutral sentiment in addition to positive and negative could also be taken into consideration, as demonstrated in Salameh et al. (2015). Beyond sentiment polarity, the text could be analysed at a finer-grained level to detect emotions, which is also known as emotion analysis. This could help narrow down the broad concepts of sentiment to better capture a person's emotional state (Ahmad et al., 2020). For instance, while anger and fear express negative sentiments, each semantically represents a different emotional state. Anger is perceived as the possible driving force of collective action, whereas fear is viewed as an action inhibitor (Miller et al., 2009).

Regardless of the level of sentiment analysis, it had only been the privilege of languages with rich resources like English. Most existing studies focusing on resource-rich languages have produced extensively annotated corpora and computational tools exclusive to these languages. However, the advent of cross-lingual sentiment analysis opens up the possibility of performing sentiment analysis on resource-poor languages by leveraging the resources from richer counterparts. With cross-lingual sentiment analysis, resource-poor languages can be endowed with comparable computational ability in identifying sentiments.

Among the seven thousand languages documented across the world, only approximately 30 languages have been equipped with linguistically annotated resources (Eberhard et al., 2021; Maxwell & Hughes, 2006). While Malaysian Malay is not the most spoken language globally, it is a language that is dominantly spoken in Malaysia. Nonetheless, Malay still lacks linguistic resources for sentiment analysis, which poses a challenge in automatically identifying sentiments expressed in Malay texts on a large scale, especially on social media platforms where almost everyone shares their personal and affective experiences. There is a need for sentiment analysis in Malay to more accurately assess individual or public emotions expressed in the local language, particularly during natural disasters, pandemics or political instability in Malaysia.

Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, pages 113 - 124
May 26, 2022 ©2022 Association for Computational Linguistics

To extend the application of sentiment analysis to the Malay language, we explore transfer learning through cross-lingual word embeddings. We also refine the cross-lingual word embeddings to capture the sentiment relationship between two languages. In this study, we use English as the source language and (Malaysian) Malay as the target language. We build a hierarchical attention model that is pre-trained using annotated English tweets and fine-tuned on a small number of Malay tweets using refined cross-lingual word embeddings. By employing such an approach, we show that the word-level attention on English tweets can be transferred to Malay tweets. In other words, if certain English words carry more weight in expressing the underlying emotions of the tweets, the corresponding Malay words sharing similar sentiment meaning would also carry the same amount of emotional weight in Malay tweets.

To the best of our knowledge, there is no gold-standard Malay emotion corpus available for our emotion classification task. The publicly available Malay tweet corpus (Husein, 2018) was previously annotated with emotions using a rule-based classifier. The rule-based classifier that relied on lexicon matching to assign emotions was not able to capture the overall context of the tweets and thus was likely to assign inaccurate emotions. We subset Husein (2018)'s corpus randomly and provide additional validation to create our gold standard Malay emotion corpus. Additionally, we also attempt to recover the truncated part of incomplete tweets to make them contextually complete.

The contributions of this study are three-fold: **a)** We demonstrate the feasibility of training the model using only English tweets to classify emotions from Malay tweets, unlike previous studies, which relied on machine translation to produce parallel training corpora to train hierarchical attention models. **b)** Our results can be used as the benchmark for any future studies as this is the first study to explore cross-lingual emotion classification in the Malay language. **c)** We validate and create a gold standard Malay emotion corpus that can be used to advance future research in the Malay language.

2 Related Work

The main challenge in cross-lingual sentiment analysis is how to bridge the language gap between the source (rich-resourced) and target (low-resourced) languages. The approaches adopted by prior studies had the element of machine translation of varying degrees.

One approach uses direct translation. Wan (2012) translated Chinese reviews into English and classified the translated English reviews using a rule-based classifier or support vector machine. Salameh et al. (2015) translated Arabic social media posts to English using their in-house machine translation system and manual translation. The Arabic-to-English translated posts were then classified automatically and manually.

Another approach tries to project annotations from the source language to the target language. Mihalcea et al. (2007) annotated the English side of the English-Romanian parallel corpus automatically using a rule-based classifier and a Naïve Bayes classifier from OpinionFinder before projecting the annotations to the Romanian-side corpus for training. Balahur and Turchi (2014) translated English sentences into French, German and Spanish, and the English-side annotations were then projected to their corresponding translated sentences to train a classifier respectively.

The third approach uses joint learning. Banea et al. (2010) translated the English corpus into five different languages. They then concatenated monolingual unigram from different languages as the features to train the model. Fuadvy and Ibrahim (2019) created a synthetic multilingual training corpus by combining English movie reviews and corresponding translated Malay movie reviews. They then trained multilingual word embeddings using a blended approach that made differentiating original English and Malay words impossible. Chen et al. (2019) first translated English documents into the target language to obtain paired training documents in the training phase. They embedded every sentence in the documents with trained monolingual sentence representations and concatenated the document representations in both languages to train a classifier subsequently.

The fourth approach uses alignment. Abdalla and Hirst (2017) adopted a simple vector space transformation for which the matrix was obtained using a closed-form solution to linearly map the Spanish/Chinese monolingual word embeddings to the English vector space. These word embeddings were later used to predict the Affective Norms for English Words (ANEW) values of the word and form the sentence arrays of the reviews for sentiment classification. Ahmad et al. (2020) adopted a similar approach but constrained the transformation matrix to be orthogonal. They then pre-trained a model using the English tweets and fine-tuned it using Hindi reviews. Hassan et al. (2021) adopted MUSE (Conneau et al., 2018) to construct English-Arabic and English-Spanish cross-lingual word embeddings. The transformation matrix was first learnt using adversarial training and subsequently refined using a synthetic parallel lexicon built from the shared embeddings space. Farra (2019) constructed cross-lingual word embeddings between English and each of the 17 low-resourced languages using VecMap (Artetxe et al., 2018). The initial bilingual word pairs were constructed without supervision by exploiting similarity matrices of each language. The transformation matrix was then iteratively refined using self-learning until the convergence criterion was met. Nasharuddin et al. (2017) exploited the structural similarity of Wordnet Bahasa and English Wordnet to map synonyms in Malay to the corresponding English counterparts using synset value and POS value. To classify documents, they aggregated the polarity scores of every word in the documents, but the classification was not satisfactory. Zabha et al. (2019) also used a similar approach by considering code-switched tweets containing both English and Malay words and classified the tweets by the sign of the total sentiment score.

The fifth approach uses co-training or its variants. Wan (2009) adopted a co-training approach by translating labelled English and unlabeled Chinese reviews into the other language. Two classifiers were trained on each view of the labelled reviews, and the classifiers were retrained iteratively by augmenting their training corpus with confidently predicted unlabeled reviews. Hajmohammadi (2014) combined self-training and active learning in their study. They first trained a base classifier on English reviews to predict the unlabeled Chinese/French reviews translated to English. The unlabeled reviews predicted with high confidence and human-annotated reviews were then selected to retrain the classifier.

Two studies on the Malay language relied on lexicon-based approaches and reported less than promising results, while the third one relying on bilingual word embeddings was ambiguous. There have also not been any studies on finer-grained sentiment analysis in the Malay language. This study aims to improve cross-lingual sentiment analysis in the Malay language using a better but also less computational expensive approach at a finer-grained level on informal corpora, and our approach is similar to that by Ahmad et al. (2020).

3 Data Sources

3.1 Corpora

English training tweets are a subset of the tweets from the EmoTweet-28 corpus curated by Liew et al. (2016). Only tweets labelled with 'anger', 'fear', 'happiness', 'love', 'none (no emotion)', 'sadness' and 'surprise' were selected to match with the emotion categories available in the Malay evaluation tweets. We included only single-label tweets as the downstream task was framed as a multiclass classification problem. Table 1 shows the emotion class distribution of the English tweets. We converted every word to lowercase, removed any mentions (@username), URLs, and tags (#hashtag), converted emojis to emoticons, expanded contractions, and removed stopwords and tweets with less than three words.

Emotion	Tweet Counts
Anger	944
Fear	178
Happiness	1299
Love	385
None	7562
Sadness	349
Surprise	178

Table 1: Emotion distribution of English training tweets

Malay evaluation tweets are a random subset of the tweets available on Malaya Documentation (Husein, 2018), previously labelled using a rule-based classifier. We hired and trained three native speakers to validate the emotions using majority voting. The Malaya Documentation corpus contains both Malaysian Malay and Indonesian Malay tweets. Therefore, we adopted a hybrid approach (Google's language detector followed by human detection) to remove the Indonesian Malay tweets from our corpus. Table 2 shows the class distribution of the Malay tweets.

Emotion	Tweet Counts
Anger	304
Fear	423
Happiness	117
Love	160
None	257
Sadness	279
Surprise	366

Table 2: Emotion distribution of Malay evaluation tweets

We performed similar pre-processing steps as in the English training tweets. Contractions in Malay were first normalised and then spell-checked according to the context. For example, *msg2*[1] were expanded to *masing-masing* (individually or respectively), and *x* was expanded to *tidak* (no).

We then performed stratified sampling to select 1000 Malay evaluation tweets as the test set (**Malay test set**) and the remaining 843 tweets (after removing tweets shorter than 3 words) as the fine-tuning set (**Malay fine-tuning set**) to fine-tune our model.

3.2 Word Embeddings

Our study used the **English monolingual word embeddings (EWE)** pre-trained on tweets by Godin (2019) using the Skip-gram architecture and contained approximately 3 million words. The words were represented by 400-dimensional vectors.

Malay monolingual word embeddings (MWE) were pre-trained on tweets and Instagram posts by Husein (2018) using Skip-gram architecture and contained approximately 1.3 million words. Normalisation and spell-check were performed to standardise non-standard Malay words in these embeddings. Normalisation ensured that contractions were expanded to the full form (e.g., *x* was expanded to *tidak*). In spell-check, abbreviated words like *nnt,* which remained unchanged after normalisation, would be augmented by adding vowels, producing a list of candidate words like *nenet*, *nanto* and *nanti*. The abbreviated word would be matched to the candidate closest to a legitimate Malay word. For example, *nnt* would be corrected to *nanti* (wait or later), a legitimate Malay word, after the augmentation. This step was essential as it would ensure more word pairs to be used in the subsequent mapping as our bilingual lexicon contained standard words.

We also selected the top 800,000 most frequent words from its training corpora and compared them against the words extracted from selected corpora by *Dewan Bahasa dan Pustaka Malaysia*[2] (DBP) written in standard Malay so that non-(standard) Malay words from the vocabulary could be removed (**F-MWE**). This step minimised concurrent standard and non-standard entries of a word that could create unnecessary noise.

3.3 Bilingual Lexicon

An **English-Malay bilingual lexicon** was obtained from Malaya Documentation (Husein, 2018). Invalid words, non-English words and non-Malay words were filtered out. We randomly selected 90% of these lexicon word pairs for mapping in the training phase (**T-BL**), while the remaining 10% were used to create a set of gold standard test English-Malay word pairs. For every word pair, we retained its English side, for which we then manually extracted its corresponding Malay translations from the English-Malay dictionary by DBP to create a gold standard bilingual lexicon (**G-BL**). G-BL contains 1273 entries of which one English word can have one or many Malay translations from G-BL. G-BL consists of 3675 unique Malay words.

4 Methodology

4.1 Cross-lingual Word Embeddings

To create cross-lingual word embeddings, we mapped the English embeddings, $\boldsymbol{E}$ to the Malay embeddings space using the orthogonal transformations approach proposed by Smith et al. (2017). Malay embeddings were first made to have the same dimensions as English embeddings by post-padding with arrays of zeros. We also normalised both embeddings to a unit length.

From the bilingual lexicons (T-BL) containing n word pairs, two ordered matrices $S_D \in \mathbb{R}^{n\times 400}$ and $T_D \in \mathbb{R}^{n\times 400}$ were formed where i^{th} row of the matrices corresponded to the English and Malay word vectors of the i^{th}word pairs. We then performed Singular Value Decomposition (SVD) operation on the matrix product $P = {S_D}^T T_D \in$

[1] It is common in non-standard Malay to form contraction indicating reduplication using a number suffix based on how many times the word is repeated.

[2] A government body that coordinates the use of the Malay language in Malaysia.

$\mathbb{R}^{400\times 400}$ and subsequently, P was represented by $U\sum V^T$. English embeddings, $\boldsymbol{E}$ were then aligned to the Malay embeddings space by multiplying it with the transformation matrix $\boldsymbol{O} = UV^T$ that was subject to the orthogonal constraint:

$$\max_{O} \sum_{i=1}^{n} t_i{}^T \boldsymbol{O} s_i \text{, subject to } \boldsymbol{O}^T \boldsymbol{O} = \boldsymbol{I} \tag{1}$$

4.2 Embeddings Refinement

To refine the cross-lingual word embeddings in Section 4.1, we modified Yuan et al. (2020)'s method by eliminating human intervention in capturing sentiment information. The refinement pulled words similar to the keyword closer and pushed words dissimilar to the keyword apart. We used Extended Affective Norms for English Words (E-ANEW) (Warriner et al., 2013) to determine these sentiment keywords. Words with a valence score of more than 6 (positive sentiment words) or less than 4 (negative sentiment words) were chosen.

For each keyword κ, we collected ten nearest neighbours in English and Malay languages from the cross-lingual word embeddings using cosine similarity. These nearest neighbours were then categorised to either the positive set $\mathcal{P}_\kappa$, if they were part of the WordNet synsets of the keyword or otherwise negative set $\mathcal{N}_\kappa$. To refine the neighbourhood of the keywords, we increased the similarity between the keyword and each positive word in its positive set and decreased the similarity between the keyword and each negative word in its negative set. The embeddings would be updated by minimising the following cost function:

$$C_f(\mathbf{E}) = \sum_{\kappa \in \mathrm{K}} \Big(\sum_{n \in \mathcal{N}_\kappa} \boldsymbol{E}_n^T \boldsymbol{E}_\kappa - \sum_{p \in \mathcal{P}_\kappa} \boldsymbol{E}_p^T \boldsymbol{E}_\kappa \Big) \tag{2}$$

We also preserved the topology of the embeddings by retaining the regularisation term measuring the squared Euclidean distance between the original embeddings and the refined embeddings:

$$R(\mathbf{E}) = \sum_{w \in \mathcal{V}} \| \hat{\mathbf{E}}_w - \mathbf{E}_w \|_2^2 \tag{3}$$

The final cost function is the combination of $C_f(E)$ and $R(\mathbf{E})$**:**

$$C(\mathbf{E}) = C_f(\mathbf{E}) + \lambda R(\mathbf{E}) \tag{4}$$

where we set λ to 1 in our study. Without human intervention, the categorisation of the nearest neighbours was definite and entirely dependent on the lemmas in the synsets in which most of the nearest neighbours of the keywords were categorised to the negative set. This implies that lemmas in the synsets that were semantically close to the keywords were located far apart in the embeddings space. Thus, regardless of their distance, we added lemmas that were not part of the nearest neighbours into the positive set such that they would be closer to the keywords after the refinement.

4.3 Emotions Classification Model

To classify emotions, we developed a hierarchical attention model similar to Yang et al. (2016) in which only the attention at the sentence level was swapped with a multi-head self-attention mechanism. We also experimented with swapping the original attention with a multi-head self-attention mechanism at only the word level and both word level and sentence level, but both degraded the performance significantly. The model can be divided into four main layers: the input layer, the word-level layer, the sentence-level layer and the output layer.

Input layer: For each tweet, $\boldsymbol{x}$, it contains S sentences s_i and each sentence contains W words.

Word-level layers: i) **Word encoder**: We use a BiLSTM to get the contextual information of the words from both directions. We encode the word by concatenating the hidden states from both directions. ii) **Word hidden layer**: We apply another hidden layer to encode the word annotations further to capture any complex relationship between words. iii) **Word attention**: The attention mechanism introduced by Bahdanau et al. (2016) is used to capture the weights of the words in expressing the underlying emotion in a sentence. A detailed description of the attention mechanism and how it is used to form representations can be found in Yang et al. (2016).

Sentence-level layers: i) **Sentence encoder:** We also use a BiLSTM to obtain the sentence contextual information from both directions. Similarly, we encode the sentence by concatenating the hidden states from both directions. ii) **Sentence hidden layer**: We use another hidden layer of size (64 × 1), the multiplier of the number of heads, to encode the sentence annotations further to capture any complex relationship between sentences. iii) **Sentence attention:** We swapped Bahdanau (2016)'s attention mechanisms originally used in Yang et al. (2016) with a one-head scaled dot-product attention mechanism (Vaswani et al., 2017). We set the dimension of the queries, keys and values in the attention mechanism to have the same values in this study. The encoded sentence annotations are used as the query vectors, key vectors and value vectors. To obtain the tweet representation, we apply a global max-pooling operation on the output.

Output layer: The tweet representation with dropout is then sent to the output layer. We use a hidden layer of 7 neurons to match the number of emotion classes.

4.4 Model Implementation

We performed hyperparameter tuning, pre-training, fine-tuning and evaluation for our model on Google TPU using TensorFlow 2.5.0 with Python3.

Hyperparameter tuning: The hyperparameters of the model were tuned solely on English training tweets using grid search with 5-fold cross-validation. The hyperparameters and their search space are listed in Appendix A. The optimal values are as follows: the hidden unit in the word-level hidden layer = 200, the hidden unit in the sentence-level hidden layer = 64, alphas of all Leaky ReLU functions = 0.3, dropout rate = 0.2, initial learning rate = 7e-3, epoch = 30 and batch size = 500.

Pre-training: We set the dimension for a unidirectional LSTM at both word level and sentence level to 200 dimensions and the context vectors required in the word attention to 400 dimensions. All intermediate layers were activated using Leaky ReLU. We pre-trained our model on English training tweets with frozen refined cross-lingual English embeddings, AdamW optimiser with a warm-up proportion of 0.1 and sparse categorical cross-entropy as the loss function.

Fine-tuning: All layers in the model underwent the fine-tuning process. Using the Malay fine-tuning set with our refined cross-lingual Malay embeddings, we fine-tuned our model for another 30 epochs with a default batch size of 32. The other hyperparameters and loss function remained unchanged as they were in pre-training. The optimiser's step_per_epoch was also changed accordingly.

5 Experiment Results

5.1 Bilingual Lexicon Induction

We used bilingual lexicon induction to evaluate the quality of our embeddings mapping by finding the top-10 most semantically similar Malay words to the English words in G-BL using cosine similarity from the shared vector space (P@10). P@10 measures the proportion of English words in G-BL, obtaining at least one correct translation among the 10 induced Malay translations for each English word in the G-BL. We used a more lenient measure as, unlike other studies which had embeddings trained on formal corpora, our embeddings were trained on notoriously noisy corpora. We also used this method to justify selecting the most frequent words in the embeddings' vocabulary. The results of the induction are shown in Table 3.

Embeddings	**P@10**
MWE	22.2041% (274/1234)
F-MWE	24.9167% (299/1200)

Table 3: Mapping quality between MWE and F-MWE using T-BL

Although we fixed the number of word pairs in the G-BL, F-MWE has a smaller vocabulary size and hence a different number of effective word pairs for evaluation as reflected in the denominator in P@10. The improvement in the mapping quality when using F-MWE was attributed to the reduced noise in the cross-lingual embeddings space since we had removed numerous non-(standard) Malay words from F-MWE. In other words, the English words were not obscured by irrelevant 'Malay' neighbours and could induce the correct Malay translations more easily. Although using F-MWE would not directly affect the downstream classification performance, the loading of the word embeddings was more efficient in terms of time and computational power as a large number of non-(standard) Malay words have been discarded.

Embeddings	**P@10**
MWE	24.8784% (307/1234)
F-MWE	27.3333% (328/1200)

Table 4: Mapping quality between MWE and F-MWE using N-BL

We also investigated the quality of T-BL by translating the English-side words in T-BL to Malay using Google Cloud Translation API, resulting in a new set of bilingual word pairs (**N-BL**). The results are presented in Table 4. We observed that each embedding mapping was improved approximately by about 2.5%, and this suggests that there is still room for improvement for the quality of T-BL. It is possible that the words in T-BL were paired up imprecisely. F-MWE also achieved better mapping quality than MWE, even when using N-BL. This again emphasises the importance of our filter when the embeddings were pre-trained on tweets or noisy corpora. Essentially, the embedding vectors remain unchanged for the Malay words but are significantly smaller in size.

Next, we attempted to augment N-BL using the nearest neighbours (NN) of English words in N-BL by using cosine similarity. However, realising some of the English NN were noise, we filtered out those not in Words Corpus by Natural Language Processing Toolkit (NLTK). The remaining neighbours were then translated to Malay using Google Cloud Translation API. The results of the augmentation using F-MWE are given in Table 5.

Augmentation Strategy	P@10
N-BL	27.3333% (328/1200)
N-BL + 1NN	29.6667% (356/1200)
N-BL + 5NN	32.7500% (393/1200)
N-BL + 10NN	31.8333% (382/1200)

Table 5: Mapping quality when augmenting N-BL by 1NN, 5NN and 10NN

We observed that augmentation generally led to better mapping quality as the larger set of training bilingual word pairs could cover more English/Malay words in the induction of the transformation matrix. It increased P@10 by a minimum of 2%. However, we acknowledged that having an enormous training set was not desirable, such as in the case of N-BL+10NN. It took us significantly longer than N-BL+5NN to perform the embeddings mapping, yet the performance degraded. From the results in Table 5, we decided to proceed with augmentation using 5-nearest neighbours as it yielded the best balance between translation time, training time and mapping quality in our experiment. The cross-lingual English and Malay embeddings created using EWE and F-MWE and mapped on N-BL+5NN were then used for the downstream emotion classification task.

5.2 Emotion Classification Model

We compare the performance of our model with other baselines, including a multilayer perceptron (MLP), hierarchical attention model (HAN) proposed by Yang et al. (2016), mBERT by Pires et al. (2019) and XLM-R by Conneau et al. (2020) and Malay BERT by Husein (2018).

MLP: We use a neural network of two layers. The hidden layer of 200 hidden units is activated using the Leaky-ReLU function with a default alpha value. The output layer has a Softmax activation function. The tweet representation is obtained using global average pooling. We pre-train this network using Adam optimiser with its default learning rate for 30 epochs and batch size of 500 on English training tweets. Every layer is fined-tuned on Malay fine-tuning set for another 30 epochs of batch size 32 in few shot-learning.

HAN: We modify the hierarchical attention network proposed by Yang et al. (2016) but use BiLSTM instead of BiGRU to encode tweets. Unidirectional LSTM is set to 200 dimensions. The following intermediate layers of 200 hidden units and the output layer have Leaky ReLU and Softmax with default parameters as the activation functions, respectively. The pre-training of this model in zero-shot learning and fine-tuning in few-shot learning are identical to MLP.

mBERT: We adopt the pre-trained mBERT by Pires et al. (2019) and attach an additional output layer having a SoftMax of default parameters as the activation function. The 'pooled output' representation with a dropout rate of 0.2 is fed to the output layer for classification. This model is fine-tuned using an AdamW optimiser with an initial learning rate of 3e-5 and a warm-up proportion of 0.1 for 30 epochs and a batch size of 32 on English training tweets in zero-shot learning. It is further fine-tuned on the Malay fine-tuning set using the fine-tuning setting applied to our model in few-shot learning.

XLM-R: We adopt the pre-trained XLM-RoBERTa by Conneau et al. (2020) and attach an additional output layer having a SoftMax of default parameters as the activation function. The input to the output layer and the fine-tuning processes are identical to that of mBERT in both zero-shot and few-shot learning.

Malay BERT: We adopt the monolingual tiny-BERT pre-trained by Husein (2018) and attach an additional output layer having a Softmax of default parameters as the activation function. The input to the output layer is identical to mBERT but we fine-tune the model using the Malay fine-tuning set and the settings applied to our model.

HMAN: Hierarchical multi-head attention model described in Section 4.3. The architectural difference between HAN and HMAN is that we swapped the sentence-level attention with scaled dot-product attention.

Model	Macro F1-score
MLP	0.0469
HAN	0.2890
mBERT	0.2162
XLM-R-base	0.5193
HMAN	0.2403

Table 6: Cross-lingual emotion prediction of our model and the comparison with the baselines in zero-shot learning.

Table 6 shows the performance comparison of our methods with the four baselines on zero-shot learning on the Malay test set. Although XLM-R-base yielded the best performance in zero-shot learning, our HMAN model slightly outperforms mBERT by 2.4% even when it was not exposed to the Malay language during pre-training and is significantly less computationally expensive compared to the multilingual pre-trained language models. We also experimented with more heads for sentence attention, but the model did not have significant improvement. Even though our experiment is simpler and on a different task, the results agree with that by Michel et al. (2019), claiming that most of the heads in multi-head attention are redundant in machine translation.

Model	Macro F1-score
MLP	0.7277
HAN	0.8104
mBERT	0.8925
XLM-R-base	0.9262
Malay BERT	0.8760
HMAN	0.8836

Table 7: Cross-lingual emotion prediction of our model and the comparison with selected baselines in few-shot learning.

In Table 7, we demonstrate the capability of our model after fine-tuning the model. While HAN yielded better performance on zero-shot transfer, our HMAN model outperforms it by 7.3% and is more effective after both models underwent the same fine-tuning process. HMAN's performance is at par with mBERT and is better than the monolingual Malay BERT without using considerable computational power. It is also worth mentioning that our model only falls behind XLM-R-base by 4.3 % in exchange for 6 – 7 times[3] increase in the computational speed. In fact, our model remains feasible on the CPU and can run in approximately one hour, while fine-tuning the multilingual language model takes days using the current batch size (32) and is unachievable if using the batch size (500) in our model. The fine-tuning helps in this task because it exposes our model to how a complete Malay tweet can be formed from words and sentences.

Fine-tuning Layers	Macro F1-score
Only output	0.3648
Sentence-level + Output	0.5762
All layers	0.8836

Table 8: Performance of our model HMAN on different fine-tuning layers

The performance of only fine-tuning the output layer of our model aligns with our prior expectations. As seen in Table 8, the macro F1-score drops drastically as the model does not have knowledge of how Malay words and sentences can be joined to form tweets. We also attempted to freeze only the word-level layers during fine-tuning, but the performance of the model degraded by about 30.74%. We attribute this degradation to the inability of the model in learning how Malay words are used to form sentences.

Setting	Macro F1-score	
	Zero-shot	Few-shot
With alignment	0.2403	0.8836
Without alignment	0.1379	0.8693

Table 9: Performance of our model HMAN with and without alignment in zero and few-shot learning

Table 9 compares the performance of our model with and without the word alignment. In without alignment, the monolingual English and Malay embeddings were merely combined into a single vector space without performing any English-Malay word mapping. The model degraded in both zero-shot and few-shot scenarios as expected. While it is not significant in few-shot learning, the model did not perform satisfactorily in the zero-shot scenario. Therefore, the word alignment still plays a vital role.

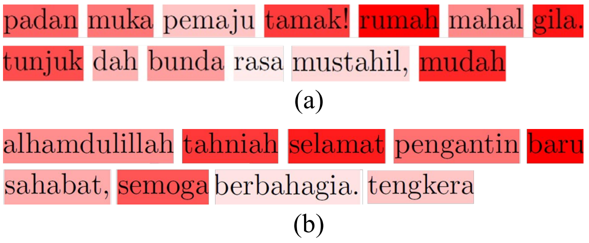
5.3 Words Attention Visualisation

To inspect how our model captures the attention of Malay words, we select two Malay tweets from the test set and visualise their attention scores using heatmaps in Figure 1. A darker shade indicates the

[3] Comparison was made on TPU using the same batch size in our model.

words receive higher attention scores, while words with a lighter shade receive lower attention scores.

We show two tweets with emotions of opposite sentiments. The tweets appear to be incomplete as we had removed stopwords in pre-processing steps. Our model can accurately place attention on the important Malay sentiment words after fine-tuning using the cross-lingual Malay embeddings.



(a)

(b)

Figure 1: Malay tweet examples with correctly predicted anger (a) and happiness (b)

In Figure 1(a), the tweeter is expressing anger at the greedy (property) developer who sells crazily expensive houses. Our model successfully places more attention on the sentiment words, *padan muka* (serve you right), *tamak* (greedy) and *gila* (crazily). *Mudah* (easy) has a darker colour here because it is treated as a sentence of only one word in our sentence tokenisation process and thus, receives all the attention score.

In Figure 1(b), the tweeter is expressing happiness and congratulating someone for getting married. The words *tahniah* (congratulation), *semoga* (wish) and the phrase *selamat pengantin baru* (happy newlyweds) were given attention correctly in the context of this tweet.

6 Conclusion and Future Work

We evaluated the quality of the existing set of Malay-English bilingual word pairs as part of the experiments in this paper and discovered that its quality could be further improved. Apart from this, we demonstrated that Malay words could benefit from their semantically and sentimentally similar English counterparts through refined cross-lingual word embeddings that were mapped using our bilingual lexicon after fine-tuning. Most importantly, our model is better than monolingual Malay BERT and at par with mBERT but utilises significantly less computational power. Even though XLM-R-base shows slightly better performance than our model by 4.3% in few-shot learning, our model is still competitive as the amount of finetuning and computational time can be reduced by 6 – 7 times. This provides us with a more cost-effective alternative to predict emotions in Malay tweets on a large scale more efficiently and possibly generalise to other languages with limited training corpora.

Unlike English, Malay remains a low-resource language with no standard Malay emotion corpus. Thus, we could not evaluate our model on other test sets to obtain a more unbiased judgement. Our Malay emotion corpus may contain some bias as the emotion labels were verified from the Malaya Documentation corpus as part of our effort to build upon existing language resources, and not annotated from scratch. Nonetheless, we hope our study can serve as the benchmark for future research, especially in English-Malay cross-lingual emotion classification using a higher quality gold-standard Malay emotion corpus we have created. Our Malay emotion corpus can be expanded in the future to include more emotion annotations. As we only performed word-level mapping and refinement, we would like to explore sentence-level mapping and refinement in future work to investigate if this will lead to further improvement. Also, we would like to evaluate our model on standard Malay emotion corpora to compare the performance of our model in formal and informal use of the Malay language.

In the future, we also plan to explore semi-supervised and unsupervised approaches such as MUSE and VecMap in creating cross-lingual word embeddings. These approaches have shown to be promising for other language pairs. Therefore, it is a possible direction to explore in building more computationally efficient cross-lingual models particularly for English-Malay that can compete with or even outperform multilingual language models.

Acknowledgement

This study was funded by "Ministry of Higher Education Malaysia for Fundamental Research Grant Scheme with Project Code: FRGS/1/2020/ICT02/USM/02/3". We would also like to thank the annotators, Siti Sakinah Ahmad Sanusi, Aqilah Syahirah Shahabudin and Nur Alya Mazlan, native speakers who helped us with the Malay corpus annotation and verification.

References

Abdalla Mohamed, and Graeme Hirst. 2017. "Cross-Lingual Sentiment Analysis without (Good)

Translation." Pp. 506–15 in Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Taipei, Taiwan: Asian Federation of Natural Language Processing.

Ahmad Zishan, Raghav Jindal, Asif Ekbal and Pushpak Bhattachharyya. 2020. "Borrow from Rich Cousin: Transfer Learning for Emotion Detection Using Cross Lingual Embedding." Expert Systems with Applications 139:112851.

Artetxe, M., Labaka, G., & Agirre, E. (2018). A Robust Self-Learning Method For Fully Unsupervised Cross-Lingual Mappings Of Word Embeddings. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 789–798.

Bahdanau Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2016. "Neural Machine Translation by Jointly Learning to Align and Translate." ArXiv:1409.0473 [Cs, Stat].

Balahur Alexandra, and Marco Turchi. 2014. "Comparative Experiments Using Supervised Learning and Machine Translation for Multilingual Sentiment Analysis." Computer Speech & Language 28(1):56–75.

Banea Carmen, Rada Mihalcea, and Janyce Wiebe. 2010. "Multilingual Subjectivity: Are More Languages Better?" Pp. 28–36 in Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010). Beijing, China: Coling 2010 Organizing Committee.

Chen Zhenpeng, Sheng Shen, Ziniu Hu, Xuan Lu, Qiaozhu Mei, and Xuanzhe Liu. 2019. "Emoji-Powered Representation Learning for Cross-Lingual Sentiment Classification." Pp. 251–62 in The World Wide Web Conference on - WWW '19. San Francisco, CA, USA: ACM Press.

Conneau Alexis, Khandelwal Kartikay, Goyal Naman, Chaudhary Vishrav, Wenzek Guillaume, Guzmán Francisco, Grave Edouard, Ott Myle, Zettlemoyer Luke and Stoyanov Veselin. 2020. Unsupervised Cross-lingual Representation Learning at Scale. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 8440–8451.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., & Jégou, H. (2018). Word Translation Without Parallel Data. ArXiv:1710.04087 [Cs].

Eberhard, M. David, F. Simons Gary, and D. Fennig Charles. 2021. Ethnologue: Languages of the World. 24th ed. Dallas: SIL International.

Farra, N. (2019). Cross-Lingual And Low-Resource Sentiment Analysis. [Doctoral Dissertation]. Columbia University.

Fuadvy Muhammad Jauharul., and Roliana Ibrahim. 2019. "Multilingual Sentiment Analysis on Social Media Disaster Data." Pp. 269–72 in 2019 International Conference on Electrical, Electronics and Information Engineering (ICEEIE). Vol. 6.

Godin, Fréderic. 2019. "Improving and Interpreting Neural Networks for Word-Level Prediction Tasks in Natural Language Processing." Ghent University, Belgium.

Hajmohammadi Mohammad Sadegh, Roliana Ibrahim, and Ali Selamat. 2014. "Density Based Active Self-Training for Cross-Lingual Sentiment Classification." Pp. 1053–59 in Advances in Computer Science and its Applications. Vol. 279, Lecture Notes in Electrical Engineering, edited by H. Y. Jeong, M. S. Obaidat, N. Y. Yen, and J. J. Park. Berlin, Heidelberg: Springer Berlin Heidelberg.

Hassan, S., Shaar, S., & Darwish, K. (2021). Cross-Lingual Emotion Detection. ArXiv:2106.06017 [Cs].

Husein Zolkepli. 2018. "Malaya, Natural-Language-Toolkit Library for Bahasa Malaysia, Powered by Deep Learning Tensorflow." Malaya.

Liew Jasy Suet Yan, Howard R. Turtle and Elizabeth D. Liddy. 2016. "EmoTweet-28: A Fine-Grained Emotion Corpus for Sentiment Analysis." Pp. 1149–56 in Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). Portorož, Slovenia: European Language Resources Association (ELRA).

Maxwell Mike and Baden Hughes. 2006. "Frontiers in Linguistic Annotation for Lower-Density Languages." Pp. 29–37 in Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006 - LAC '06. Sydney, Australia: Association for Computational Linguistics.

Meng Xinfan, Furu Wei, Xiaohua Liu, Ming Zhou, Ge Xu and Houfeng Wang. 2012. "Cross-Lingual Mixture Model for Sentiment Classification." Pp. 572–81 in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Jeju Island, Korea: Association for Computational Linguistics.

Michel Paul, Omer Levy and Graham Neubig. 2019. "Are Sixteen Heads Really Better than One?" in Advances in Neural Information Processing Systems. Vol. 32. Curran Associates, Inc.

Mihalcea Rada, Carmen Banea, and Janyce Wiebe. 2007. "Learning Multilingual Subjective Language via Cross-Lingual Projections." Pp. 976–83 in Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. Prague, Czech Republic: Association for Computational Linguistics.

Miller Daniel A., Tracey Cronin, Amber L. Garcia and Nyla R. Branscombe. 2009. "The Relative Impact of Anger and Efficacy on Collective Action Is Affected by Feelings of Fear." Group Processes & Intergroup Relations 12(4):445–62.

Nasharuddin Nurul Amelina, Muhamad Taufik Abdullah, Azreen Azman and Rabiah Abdul Kadir. 2017. "English and Malay Cross-Lingual Sentiment Lexicon Acquisition and Analysis." Pp. 467–75 in Information Science and Applications 2017. Vol. 424, Lecture Notes in Electrical Engineering, edited by K. Kim and N. Joukov. Singapore: Springer Singapore.

Pang Bo and Lillian Lee. 2005. "Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales." Pp. 115–24 in Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05). Ann Arbor, Michigan: Association for Computational Linguistics.

Pires Telmo, Schlinger Eva and Garrette, Dan. 2019. How Multilingual is Multilingual BERT? Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 4996–5001.

Salameh Mohammad, Saif Mohammad and Svetlana Kiritchenko. 2015. "Sentiment after Translation: A Case-Study on Arabic Social Media Posts." Pp. 767–77 in Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Denver, Colorado: Association for Computational Linguistics.

Smith Samuel L., David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. "Offline Bilingual Word Vectors, Orthogonal Transformations and the Inverted Softmax." ArXiv:1702.03859 [Cs].

Vaswani Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin. 2017. "Attention Is All You Need." in Advances in Neural Information Processing Systems. Vol. 30. Curran Associates, Inc.

Wan Xiaojun. 2009. "Co-Training for Cross-Lingual Sentiment Classification." Pp. 235–43 in Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Suntec, Singapore: Association for Computational Linguistics.

Wan Xiaojun. 2012. "A Comparative Study of Cross-Lingual Sentiment Classification." Pp. 24–31 in 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology. Vol. 1.

Warriner Amy Beth, Victor Kuperman and Marc Brysbaert. 2013. "Norms of Valence, Arousal, and Dominance for 13,915 English Lemmas." Behavior Research Methods 45(4):1191–1207.

Yang Zichao, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. "Hierarchical Attention Networks for Document Classification." Pp. 1480–89 in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California: Association for Computational Linguistics.

Yuan Michelle, Mozhi Zhang, Benjamin Van Durme, Leah Findlater and Jordan Boyd-Graber. 2020. "Interactive Refinement of Cross-Lingual Word Embeddings." Pp. 5984–96 in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics.

Zabha Nur Imanina, Zakiah Ayop, Syarulnaziah Anawar, Erman Hamid and Zaheera Zainal. 2019. "Developing Cross-Lingual Sentiment Analysis of Malay Twitter Data Using Lexicon-Based Approach." International Journal of Advanced Computer Science and Applications 10(1).

A. Appendix A: Hyperparameter Search Space

Some hyperparameters would be fixed throughout the experiments, such as the number of units in the unidirectional word-level/sentence-level LSTM layer and the number of heads in the sentence self-attention. The search space of the hyperparameters is as below:

Hyperparameters	**Search Space**
Number of Units in Word Hidden Layer	[100,200,300,400,500,600,700,800,900,1000]
Alpha for Word-level Leaky ReLU	[0.01,0.02,0.03,0.04,0.05, 0.3, 0.4, 0.5]
Number of Units in Sentence Hidden Layer	64
Alpha for Sentence-level Leaky ReLU	[0.01,0.02,0.03,0.04,0.05, 0.3, 0.4, 0.5]
Dropout Rate before Output Layer	[0.1,0.2,0.3,0.4,0.5]
Initial Learning Rate of AdamW Optimizer	[0.001,0.002,0.003,0.004,0.005,0.006,0.007,0.008,0.009]
Epoch	[10,20,30,40]
Batch Size	[500,600,700,800,900,1000]