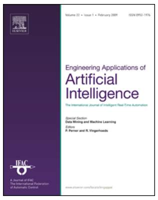
Engineering Applications of Artificial Intelligence 36 (2014) 195–203

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Cross-lingual sentiment classification using multiple source languages in multi-view semi-supervised learning

Mohammad Sadegh Hajmohammadi, Roliana Ibrahim*, Ali Selamat

Software Engineering Research Group, Faculty of Computing, Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor, Malaysia

ARTICLE INFO

Article history:
Received 5 April 2014
Received in revised form
15 July 2014
Accepted 25 July 2014
Available online 24 August 2014

Keywords:
Cross-lingual
Sentiment classification
Multiple source languages
Multi-view learning
Semi-supervised learning

ABSTRACT

Cross-lingual sentiment classification aims to utilize annotated sentiment resources in one language (typically English) for sentiment classification of text documents in another language. Most existing research works rely on automatic machine translation services to directly project information from one language to another. However, due to the existence of differing linguistic terms and writing styles between different languages, translated data cannot cover all vocabularies which exist in the original data. Further, different term distribution between translated data and original data can lead to low performance in cross-lingual sentiment classification. To overcome these problems, we propose a new model which uses labelled data from multiple source languages in a multi-view semi-supervised learning approach so as to incorporate unlabelled data from the target language into the learning process. The proposed model was applied to book review datasets in four different languages. Experiments have shown that our model can effectively improve the cross-lingual sentiment classification performance in comparison with some baseline methods.

1. Introduction

Along with the rapid increase of internet access in the world today, the volume of user-generated content on the web has also intensified. Due to the enormous amount of this content, the task of summarizing their information into a useful format is a very difficult and challenging problem. This challenge has motivated the natural language processing (NLP) communities to design and develop computational methods by which to mine and analyze the information of these text documents. Opinion mining or sentiment analysis is one of the most interesting fields in this area; it analyzes people's opinions, attitudes and sentiments towards entities such as products, individuals, events, etc. (Liu, 2012). Text sentiment classification refers to the task of determining the sentiment polarity (e.g. positive or negative) of a given text document and has received considerable attention due to its many useful applications in product review classification (Zhou et al., 2013) and opinion summarization (Ku et al., 2006).

Up until now, different methods have been used in sentiment classification. These methods can be categorized into two main groups, namely: lexicon-based and corpus-based. The lexicon-based methods classify text documents based on the polarity of words and phrases contained in the text (Taboada et al., 2011; Turney, 2002). This group of methods requires a sentiment lexicon to distinguish between positive and negative terms. In contrast, corpus-based methods train a sentiment classifier based on a labelled corpus using machine learning classification algorithms (Moraes et al., 2013; Pang et al., 2002). The performance of these methods depends intensively on both the quantity and quality of labelled corpus items as the training set.

Sentiment lexicons and annotated sentiment corpora are the most important resources for sentiment classification. However, since most recent research studies in sentiment classification are written in a limited number of languages (always English), this has led to a scarcity of labelled corpus and sentiment lexicons in other languages (Martín-Valdivia et al., 2013; Wan, 2011). Further, manual construction of reliable sentiment resources is a very difficult and time-consuming task. Therefore, the challenge is how to utilize labelled sentiment resources in one language (i.e. English) for sentiment classification into another language. This leads to an interesting research area called cross-lingual sentiment classification (CLSC). The most direct solution to this problem is the use of machine translation systems to directly project the information of data from one language into another (Balahur et al., 2014; Banea et al., 2008; Martín-Valdivia et al., 2013; Prettenhofer and Stein, 2010; Wan, 2009, 2011). Most existing works in this area have used machine translation systems to translate labelled training data from the source language into the target language and perform sentiment classification into the target language (Balahur and Turchi, 2014; Banea et al., 2010). Some other

* Corresponding author. Tel.: +60 7 5538727.
E-mail address: roliana@utm.my (R. Ibrahim).

http://dx.doi.org/10.1016/j.engappai.2014.07.020
0952-1976/© 2014 Elsevier Ltd. All rights reserved.

researchers have employed machine translation in the opposite direction to translate unlabelled test data from the target language into the source language and to perform the classification in the source language (Hajmohammadi et al., 2014b; Martín-Valdivia et al., 2013; Prettenhofer and Stein, 2010). A limited number of research works have used both directions of translation to create two different views of the training and test data to compensate for some of the translation limitations (Hajmohammadi et al., 2014a; Pan et al., 2011; Wan, 2009, 2011).

However, because the training set and the test set come from two different languages having differing linguistic terms and writing styles, as well as originating from different cultures, translated text documents cannot cover all the vocabularies contained in the original text documents. Therefore, these methods cannot attain the performance results of monolingual sentiment classification methods in which the training and test samples are from the same language. Using multiple resources from multiple languages can alleviate the problem of vocabulary coverage in CLSC. This occurs because some vocabularies which are not covered by the feature set extracted from translated documents of the one source language may be covered by the feature set of another source language. This means that feature sets extracted from the training data of multiple source languages can cover more vocabularies of test documents in the target language. For example, the translation of “awesome” into French is “génial” but a word in German with the same meaning “fantastisch” is translated to “fantastique” in French. Both words “génial” and “fantastique” are used in the French reviews and each word is covered by a different source language. Therefore, using a multiple source language technique is expected to show better performance in CLSC when compared with models which use only one source language.

Different term distribution between the original and the translated text documents is another important factor that can reduce the performance of CLSC. It means that a term may be frequently used in one language to express an opinion while the translation of that term is rarely used in the other language. To overcome this problem, making use of unlabelled data from the target language can be helpful, since this type of data is always easy to obtain and has the same term distribution as the test data. Therefore, employing unlabelled data from the target language in the learning process is expected to result in better classification in CLSC.

In the light of difficulties for CLSC, we address the task of CLSC via a multi-view semi-supervised learning framework. Specifically, we propose a new learning model that uses labelled data from multiple languages (in this paper, two languages) as multiple training datasets. Both directions of translation are then used to create different views of data. These individual views are then employed in a multi-view semi-supervised learning process to incorporate unlabelled data from the target language in the learning process.

The contributions of our work are as follows: (1) utilizing training data and their translations from multiple source languages in CLSC to cover more vocabularies of test documents in the target language; (2) employing a multi-view semi-supervised learning strategy in order to incorporate unlabelled examples from the target language in the learning process of CLSC. This is achieved by creating multiple views from the documents in both the source and the target languages through automatic machine translation and using the “majority teaching minority” strategy to select the most confident pseudo-labelled examples from unlabelled documents and adding them to the training sets in each of the individual views.

The proposed model was applied to book review datasets in four different languages. Experiments showed that the use of this model obtained better performance in comparison with some baseline methods.

The reminder of this paper is organized as follows: the next section presents related works on CLSC. Section 3 describes multiple views data creation. The proposed model is described in Section 4, while an evaluation is given in Section 5. Finally, Section 6 concludes this paper and outlines ideas for future research.

2. Related works

Cross-lingual sentiment classification has been extensively studied in recent years. These research studies are based on the use of annotated data in the source language (always English) to compensate for the lack of labelled data in the target language. Most approaches focus on resource adaptation from one language to another with few sentiment resources. For example, Mihalcea et al. (2007) generate subjectivity analysis resources into a new language from English sentiment resources by using a bilingual dictionary. In other works (Banea et al., 2010; Banea et al., 2008), automatic machine translation engines were used to translate the English resources for subjectivity analysis. In a further study (Banea et al., 2008), the authors showed that automatic machine translation is a viable alternative to the construction of resources for subjectivity analysis in a new language. In two different experiments, they first translated training data of subjectivity classification from the source language into the target language. They then utilized this translated data to train a classifier in the target language and applied this trained classifier to classify test data. Additionally, in another experiment, machine translation was used to translate test data from the target language into the source language and a classifier was then trained based on training data in the source language. After the training phase, the translated test data was presented to the classifier for sentiment polarity prediction. Wan (2008) used unsupervised sentiment polarity classification in Chinese product reviews. He translated Chinese reviews into different English reviews using a variety of machine translation engines and then performed sentiment analysis for both the Chinese and English reviews using the lexicon-based technique. Finally, he used ensemble methods by which to combine the analysis results. This method requires sentiment lexicon in the target language and cannot be applied to other languages with no lexicon resource. Pan et al. (2011) designed a bi-view non-negative matrix tri-factorization (BNMTF) model in an attempt to solve the problem of cross-lingual sentiment classification. They used the machine translation to achieve two representations of training and test data; one in the source language and another in the target language. This model was then used to combine the information from two views.

Another approach is that of feature translation, which involves translating the features extracted from labelled documents (Moh and Zhang, 2012; Shi et al., 2010). The features, selected by a feature selection algorithm, are translated into different languages. Subsequently, based on those translated features, a new model is trained for each language. This approach only needs a bilingual dictionary to translate the selected features. It can, however, suffer from the inaccuracies of dictionary translation, in that words may have different meanings in different contexts. Therefore, selecting the features to be translated can be an intricate process. Prettenhofer and Stein (2010) investigated CLSC from the domain adaptation view by employing structural correspondence learning (SCL) (Blitzer et al., 2006). They adapted SCL to use unlabelled data and a word translation oracle to induce correspondence among the words from both the source and target languages. They first selected some word pairs called pivots and then identified correlations between pivots and other words in unlabelled documents. After that, a map was extracted to associate the original representation of a document in the source and target languages with its cross-lingual representation. The classification was performed in the new mapped space.

In another work, Wan (2009, 2011) used the co-training algorithm to overcome the problem of CLSC. He first investigated basic methods for CLSC by using machine translation services. He then exploited a bilingual co-training approach to leverage annotated English resources to sentiment classification in Chinese reviews. In this work, firstly, machine translation services were used to translate English labelled documents (training documents) into Chinese and similarly, Chinese unlabelled documents into English. The author used two different views (English and Chinese) in order to exploit the co-training approach into the classification problem. Co-training usually selects high confidence examples to add to the training data. If, however, the initial classifiers in each view are not good enough, the probability of adding examples having incorrect labels to the training set will be increased. Therefore, adding "noisy" examples not only cannot increase the accuracy of the learning model, but will also gradually decrease the performance of each classifier.

3. Multiple views creation

The first step in the construction of a multi-view semi-supervised learning model is the creation of different views of data. For this task, labelled training examples from each of the source languages are translated into the target language and combined to create training data in the target language. In another instance, unlabelled examples are translated from the target language into each of the source languages. Therefore, we have labelled and unlabelled examples in both the source and target languages. In this paper, we used labelled data from two different languages as the source languages. Consequently, we have both labelled and unlabelled examples in three different views, specifically: source language 1, source language 2 and target language. Fig. 1 shows the process of multi-view data creation. We sorted these views into, namely: view1, view2 and view3 as shown in Fig. 1. In each view, there is an individual feature set that was extracted from the training data of the corresponding view. Therefore, both labelled and unlabelled examples are represented based on the corresponding feature set of each view.

4. Multiple source languages multi-view (MLMV) semi-supervised learning model

The aim of the approach proposed in this paper is to improve the performance of CLSC by incorporating unlabelled data from the target language into a multi-view semi-supervised learning model. This is achieved by using labelled data from multiple source languages. In this model, each unlabelled example is classified from different views and based on different feature sets. Those unlabelled examples that are confidently classified by majority views are selected to add to the training set of minor views in an iterative process. This means that if major classifiers in different views are confident and in agreement with the predicted label of an unlabelled example, this example can be added to the training set of other views which disagree with major views. Confidence of label prediction for major views is calculated by averaging the individual confidence of classifiers in agreeing views. This strategy is similar to the "majority teaching minority" strategy introduced in a previous study (Zhou and Li, 2010). In the case of three classifiers; if two classifiers agree on the predicted label of a set of unlabelled examples, the examples from the set having maximum average confidence will be selected as the most confident examples. These will then be added to the training set of another classifier. Average confidence of an example is calculated by averaging the confidence of majority classifiers in predicting the label of that example. The framework of the proposed model is illustrated in Fig. 2. The detailed procedure is shown as follows:

Algorithm 1. Multiple source Languages Multi-View (MLMV) learning

```
Given: $L_1$: The initial labelled training set from source language 1
       $L_2$: The initial labelled training set from source language 2
       $U$: The pool of unlabelled examples from the target language
       $V_1$, $V_2$ and $V_3$: Three different views of data
Initial parameters: $p$: number of most confident positive examples selected by each view in each cycle
                    $n$: number of most confident negative examples selected by each view in each cycle
 –Train classifier $h_1$ on view $V_1$ of $L_1$.
 –Train classifier $h_2$ on view $V_2$ of $L_2$.
 –Train classifier $h_3$ on view $V_3$ of $L_3 = L_1 + L_2$.
 –Loop for predefined number of iterations
    –Use $h_1$, $h_2$ and $h_3$ to predict class label and calculate the prediction confidence of each example in $U$
    –For $i = 1..3$
        –Let $P_i$ set of unlabelled examples that $h_k$ and $h_j$ ($j, k \neq i$) agree on their predicted labels.
        –Calculate average confidence for each of the examples in $P_i$ by averaging the prediction confidence of $h_j$ and $h_k$.
        –Select $p$ positive and $n$ negative examples having the highest average confidence from $P_i$.
    –Add selected examples to the $L_i$ and remove them from $U$.
    –End for
    –Retrain classifier $h_1$ on view $V_1$ of new $L_1$.
    –Retrain classifier $h_2$ on view $V_2$ of new $L_2$.
    Retrain classifier $h_3$ on view $V_3$ of new $L_3$.
 –End loop
```

In this algorithm, after creating different views of data, three individual classifiers (h_1, h_2 and h_3) are trained based on the

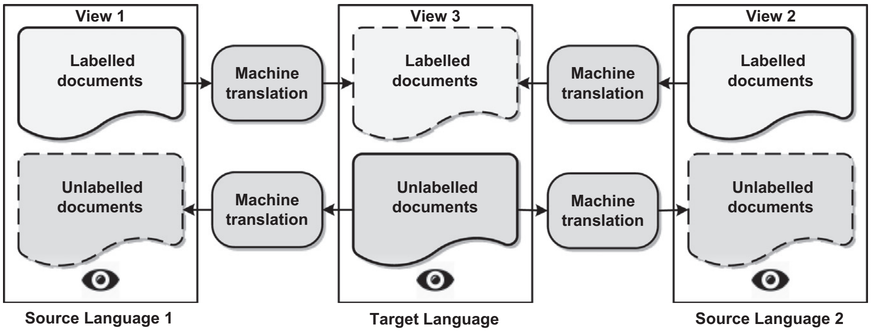


Fig. 1. Multi-view data creation.

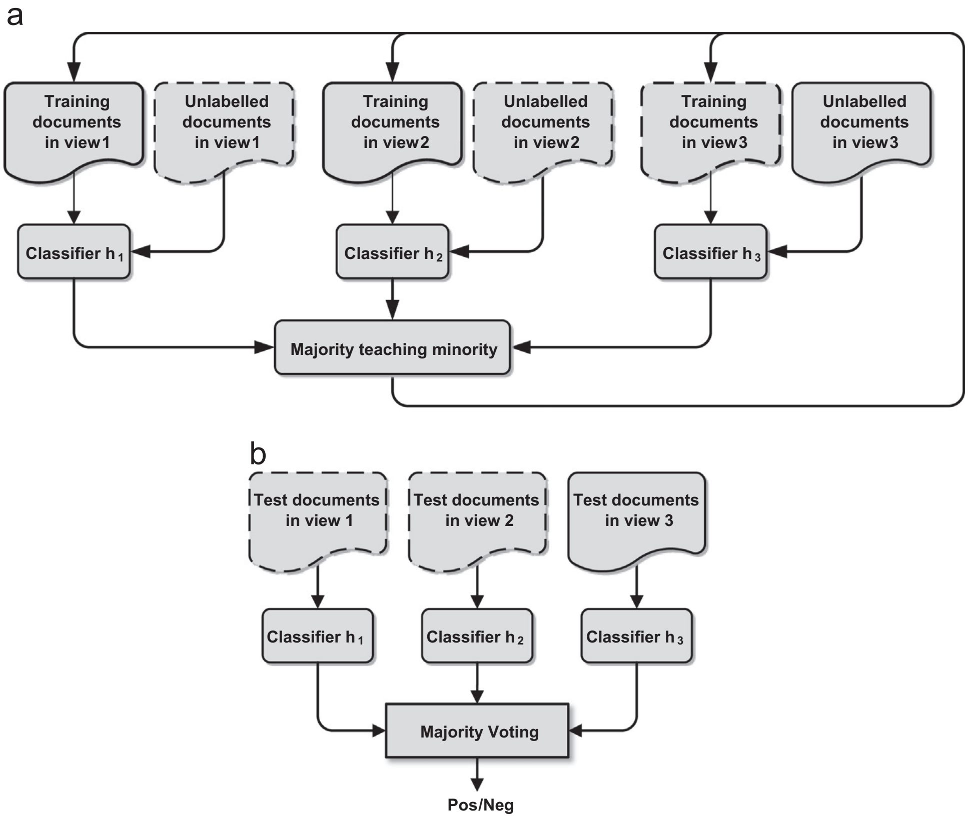

a
Training documents in view1
Unlabelled documents in view1
Training documents in view2
Unlabelled documents in view2
Training documents in view3
Unlabelled documents in view3
Classifier h1
Classifier h2
Classifier h3
Majority teaching minority
b
Test documents in view 1
Test documents in view 2
Test documents in view 3
Classifier h1
Classifier h2
Classifier h3
Majority Voting
Pos/Neg

Fig. 2. Framework of the proposed model. (a) Learning phase and (b) test phase.

training set of each view. They are then applied to the unlabelled examples pool in the corresponding view. Following this, in each view, a set of unlabelled examples which classifiers in other views agree with the prediction of their labels are identified and the most confident p positive and n negative examples from this set are then selected as pseudo-labelled examples and added to the corresponding training set. The prediction confidences for these examples are calculated based on averaging the confidence of agreeing classifiers. The process of training set enrichment is repeated for a predefined number of iterations. After the full learning process, the independent test samples are presented to the trained classifiers based on corresponding views. The final prediction label for each test example is then computed based on the majority voting approach. The process of creating different views for test examples is similar to the process for unlabelled examples as explained in Section 3.

5. Evaluation

In this section, we evaluate our proposed model for cross-lingual sentiment classification in four different languages in the book review domain and compare it with some baseline methods.

5.1. Datasets

We have selected book review documents from two different cross-lingual sentiment datasets. The first dataset was used by (Prettenhofer and Stein, 2010), and contains Amazon product reviews for three different domains consisting of books, DVDs and music. These are in four different languages, specifically: English, French, German and Japanese. Each review document is labelled as being either positive or negative based on its sentiment polarity. We only selected book reviews from this dataset. The book review dataset in the English language contains 2000 (1000 positive and 1000 negative) documents considered as being one of the source language data. A total of 6000 review documents (3000 positive and 3000 negative) were selected from each of the French, German and Japanese languages respectively and considered as being target languages data. Another dataset used in this paper is the pan reviews dataset (Pan et al., 2011). This collection consists of three review datasets in different domains (movie, book and music) in both English and Chinese. We selected only book reviews in Chinese from this collection. This dataset also contains 4000 book review documents (2000 positive and 2000 negative) and is considered as unlabelled data in the Chinese language.

By combining review documents from these two datasets, four different evaluation datasets for cross-lingual sentiment classification were consequently formed as follows:

1. EnGe-Fr: In this set, French is considered as the target language while English and German are used as two different source languages.
2. EnFr-Ge: In this set, German is considered as the target language while English and French are used as two different source languages.
3. EnFr-Jp: In this set, Japanese is considered as the target language while English and French are used as two different source languages.
4. EnJp-Ch: In this set, Chinese is considered as the target language while English and Japanese are used as two different source languages.

In all datasets, all reviews in the source languages are translated into target languages and similarly, all reviews in target languages are translated into the source languages using the Google Translate engine (http://translate.google.com/). Table 1 shows the properties of these four evaluation datasets.

Table 1
Details of the datasets used in our experiments.

Dataset	Domain	Languages		Total documents	Positive documents	Negative documents	Reference
EnGe-Fr	Book review	Source Language 1	English	2000	1000	1000	Prettenhofer and Stein, (2010)
		Source Language 2	German	2000	1000	1000	
		Target Language	French	6000	3000	3000	
EnFr-Ge	Book review	Source Language 1	English	2000	1000	1000	
		Source Language 2	French	2000	1000	1000	
		Target Language	German	6000	3000	3000	
EnFr-Jp	Book review	Source Language 1	English	2000	1000	1000	
		Source Language 2	French	2000	1000	1000	
		Target Language	Japanese	6000	3000	3000	
EnJp-Ch	Book review	Source Language 1	English	2000	1000	1000	
		Source Language 2	Japanese	2000	1000	1000	
		Target Language	Chinese	4000	2000	2000	Pan et al., (2011)

In the pre-processing step, all English, French and German reviews are converted into lowercase. Special symbols and other unnecessary characters are eliminated from each review document. In the Japanese text document, we applied MeCab[1] segmenter software to segment the reviews; while Chinese documents were segmented by the Stanford Chinese word segmenter.[2] In the feature extraction step, unigram and bi-gram patterns were extracted as sentimental patterns. To reduce computational complexity, we performed feature selection using the information gain (IG) technique. We selected 5000 high score unigrams and bi-grams as final features. Each document is represented by a feature vector. Each entry of a feature vector contains a feature weight. We used term presence as feature weights because this method has been confirmed as the most effective feature-weighting method in sentiment classification (Pang et al., 2002; Xia et al., 2011).

5.2. Baseline methods

The following baseline methods were implemented in order to compare the effectiveness of proposed models using the same system. Based on existing literature, the co-training (Wan, 2009, 2011) and the structural correspondence learning (SCL) (Prettenhofer and Stein, 2010) algorithms are two of the most well-known and best-performing methods that previously applied to the CLSC.

1. Multiple source Language Single View learning model (MLSV): This model also uses labelled data from two different source languages as training data but only one direction of translation is used. In this model, unlabelled documents are translated from the target language into two source languages. A traditional co-training algorithm (Blum and Mitchell, 1998) is used with different training datasets in each language (co-training in view1 and view2). This model is then compared to the proposed model in order to evaluate the effect of using bidirectional translation to create multiple views in semi-supervised learning.
2. Single source Language Multiple View learning model in the first Source Language (SLMV-S1): This is the traditional co-training algorithm which was used in the study by (Wan, 2009, 2011). It used labelled data from the first language as the source language data and unlabelled data from the target language in two views (view1 and view3).
3. Single source Language Multiple View learning model in the second Source Language (SLMV-S2): This is the traditional co-training algorithm which was used in a paper by (Wan, 2009, 2011). It used labelled data from the second language as source language data and unlabelled data from the target language in two views (view2 and view3).
4. Ensemble of SLMV-S1 and SLMV-S2 models (SLMV-S12): This is the combination of the SLMV-S1 and SLMV-S2 model. The output prediction of this model is calculated based on the average of each individual model.
5. Structural Correspondence Learning model (SCL): We implemented this model as introduced in (Prettenhofer and Stein, 2010). We used the Google Translate service to map words in the source vocabulary to the corresponding translation in the target vocabulary. Other parameters are set as used in (Prettenhofer and Stein, 2010). This method is implemented only in the case of using English as the source language.

5.3. Experiment setup

In all experiments, we used the support vector machine classifier (SVM) (Joachims, 1999) as the base classifier in each view in all semi-supervised methods. SVM^{light} (http://svmlight.joachims.org/) is used as the SVM classifier in the experiments with all parameters set to their default values. The output value of the SVM classifier for a review document indicates the confidence level of its label prediction. The sign of the prediction value indicates the sentiment polarity of a document.

5.3.1. Cross-validation in semi-supervised learning

To generate reliable results, we performed a 3-fold cross validation on semi-supervised learning. For this task, the unlabelled documents in the target language are randomly divided into three groups of equal size. In each step of the cross validation, two groups of documents are treated as the unlabelled pool and the evaluation of the performance is based on the remaining group as an independent test set. The final results are averaged over three iterations.

5.3.2. Performance measure

Generally, the performance of sentiment classification is evaluated by using four indexes, namely: Accuracy, Precision, Recall and F1-score. Accuracy is the proportion of all true predicted instances against all predicted instances. An accuracy of 100% means that the predicted instances are exactly the same as the actual instances. Precision refers to the portion of true predicted instances against all predicted instances for each class. Recall denotes the portion of true predicted instances against all actual instances for each class. F1 is a harmonic average of precision and recall.

[1] http://mecab.googlecode.com/svn/trunk/mecab/
[2] http://nlp.stanford.edu/software/segmenter.shtml

Table 2
Performance comparison of four datasets after completion of full learning process (best results are reported in bold-face type).

Dataset	Methods	Accuracy	Positive			Negative		
			Precision	Recall	F1	Precision	Recall	F1
EnGe-Fr	MLMV	**79.85**	**78.81**	81.67	**80.20**	81.00	**78.03**	**79.48**
	MLSV	78.92	75.83	**84.93**	80.12	**82.87**	72.90	77.56
	SLMV-S1	77.52	75.45	81.63	78.39	80.06	73.40	76.55
	SLMV-S2	77.97	75.28	83.27	79.07	81.30	72.67	76.74
	SLMV-S12	78.83	76.42	83.40	79.75	81.76	74.27	77.82
	SCL	78.41	76.00	82.98	79.33	81.34	73.84	77.40
EnFr-Ge	MLMV	**81.55**	83.60	78.50	80.97	79.74	84.60	**82.10**
	MLSV	80.47	81.14	79.40	80.25	79.85	81.53	80.67
	SLMV-S1	80.03	79.54	80.87	80.20	80.54	79.20	79.86
	SLMV-S2	79.17	79.21	79.10	79.15	79.13	79.23	79.18
	SLMV-S12	81.03	80.41	**82.07**	**81.23**	**81.69**	80.00	80.84
	SCL	79.01	**83.82**	72.16	77.56	75.37	**85.86**	80.27
EnFr-Jp	MLMV	**73.73**	76.97	67.97	72.15	71.26	79.50	**75.13**
	MLSV	71.68	73.48	67.97	70.60	70.16	75.40	72.68
	SLMV-S1	72.27	74.54	67.73	70.96	70.40	76.80	73.45
	SLMV-S2	72.45	74.08	**69.13**	71.51	71.04	75.77	73.32
	SLMV-S12	73.57	75.99	68.90	**72.27**	**71.56**	78.23	74.75
	SCL	72.90	**77.05**	65.30	70.67	69.87	**80.50**	74.80
EnJp-Ch	MLMV	**76.65**	77.82	**74.60**	**76.16**	**75.60**	78.70	77.11
	MLSV	74.37	80.68	64.10	71.43	70.23	84.65	76.76
	SLMV-S1	75.32	79.17	68.75	73.59	72.38	81.90	76.84
	SLMV-S2	70.50	80.56	54.15	64.74	65.44	86.85	74.63
	SLMV-S12	75.12	**83.59**	62.60	71.56	70.10	**87.65**	**77.89**
	SCL	70.58	70.89	69.24	70.06	70.28	71.90	71.08

5.4. Results and discussion

In this section, our proposed method is compared with five baseline methods. We used $p=n=5$ for our proposed model and all co-training algorithms as in (Wan, 2011). The total number of iterations is set to 30 iterations for all iterative algorithms. Table 2 shows the comparison results after the full learning process. As we can see in this table, the proposed model outperforms all baseline methods, especially regarding accuracy in all datasets. These results show that the use of multiple source languages in a multi-view learning approach can improve the accuracy of CLSC. By comparing the MLMV and MLSV models, we can conclude that using training and test documents in two different (original and translated) forms in both the source and target languages has a beneficial effect on classification performance. This can be attributed to the fact that the learning model uses original documents in at least one learning component of the model and consequently alleviates the destructive effects of translation errors.

Compared to the SLMV-S1 and SLMV-S2 models, MLMV shows better overall accuracy in all datasets. Due to the use of training data from more languages, more vocabularies can be covered from documents in the target language. Consequently, the classification accuracy is improved in comparison to the single source language models. Therefore, it can be concluded that using multiple source languages has a beneficial effect on the performance of CLSC.

As shown in this table, SLMV-S2 indicates that the worst performance occurred in the EnJp-Ch dataset. This means that cross-lingual sentiment classification in the Chinese language using Japanese labelled documents cannot result in a reliable outcome. However, in spite of low performance in Japanese-Chinese cross-lingual sentiment classification, the MLMV model outperforms all baseline methods in this dataset, at least in terms of accuracy. These results support the idea that the combination of multi-views in a multiple source language framework can help to improve the performance of CLSC.

In order to assess whether there are any significant differences in terms of accuracy between the proposed model and semi-supervised baseline methods, we conducted a statistical test based on accuracy results obtained from 3-fold cross-validation. We used a paired t-test to evaluate whether differences between two methods are statistically significant. Table 3 shows the numerical results of the statistical test. With the exception of those between MLMV and SLMV-S12 in the EnFr-Jp and EnJp-Ch datasets, all other comparisons showed statistically significant differences, for a significant level of $\alpha=0.05$.

Fig. 3 shows the learning curves of various methods on four evaluation datasets. Each fold of cross-validation generated a learning curve for the experiment of each model. The final learning curve was determined using the average accuracies of each point from generated curves. In the first two datasets, the proposed model shows the best accuracy from among all baseline methods during the learning process. In the EnFr-Jp and EnJp-Ch datasets, at the starting point of the learning process, the ensemble of co-training model (SLMV-S12) shows better performance in comparison to the proposed model. However, after some learning cycles, the accuracy of the proposed model overtakes all baseline methods.

Fig. 4 compares the accuracy of combined views and each of the individual views of the proposed model during the learning process in four datasets. This figure shows that all individual views improved during the learning process. This means that the strategy of “majority teaching minority” assists in the improvement of each of the views. This figure also demonstrates that the combination of three views in CLSC outperforms all individual views. This supports the idea that the information in multiple views can complement each other to cover more vocabularies in the target language. Each view can cover some limited terms of test examples and a combination of these views covers more terms in the test examples.

6. Conclusion and future work

This paper has proposed an MLMV learning model for CLSC. It creates multiple views of both labelled and unlabelled documents by incorporating multiple source languages and an automatic

Table 3
The *p*-value of paired *t*-test that compares MLMV model with baseline methods for each dataset ("Y": statistically significant; "N": Not statistically significant).

Dataset	Methods	*P*-Value	Significant?
EnGe-Fr	MLSV	3.4356E-05	Y
	SLMV-S1	2.7403E-09	Y
	SLMV-S2	2.5257E-05	Y
	SLMV-S12	3.5621E-04	Y
EnFr-Ge	MLSV	3.7619E-04	Y
	SLMV-S1	1.4596E-08	Y
	SLMV-S2	6.7627E-08	Y
	SLMV-S12	2.9871E-03	Y
EnFr-Jp	MLSV	5.3498E-03	Y
	SLMV-S1	3.7891E-02	Y
	SLMV-S2	2.1596E-03	Y
	SLMV-S12	6.5309E-01	N
EnJp-Ch	MLSV	4.7541E-06	Y
	SLMV-S1	1.6725E-02	Y
	SLMV-S2	1.9536E-08	Y
	SLMV-S12	6.7740E-01	N

machine translation engine. A multi-view semi-supervised learning strategy with "majority teaching minority" focus has been used to incorporate unlabelled examples from the target language in the learning process to improve the performance of CLSC. We conducted experiments on different datasets from different languages. Our proposed model was evaluated by comparing its performance to the performance of some baseline methods. The experimental results have shown that using multiple source languages in a multi-view framework can help to improve the performance of CLSC. In fact, different views and different source languages can complement each other to cover the sentimental terms of test data. As a result, better performance in sentiment classification is achieved.

However, the selection of an appropriate language to be used as the source language proved to be a challenging task in this model. The similarity of source language and target language in terms of linguistic expression and writing style can help the cross-lingual sentiment classification. On the other hand, our proposed method obviously depends on the availability of the automatic machine translation engines for the source and the target languages respectively. Although most of the commercial machine translation engines have the capability of translating text documents from and into a large number of languages, they cannot be used to freely translate large amounts of data. This issue may limit the use of machine translation in our proposed method.

In a future work, we plan to exploit different learning models via a combination of different views and multiple source languages with regard to the learning process. As a result of this, each learning

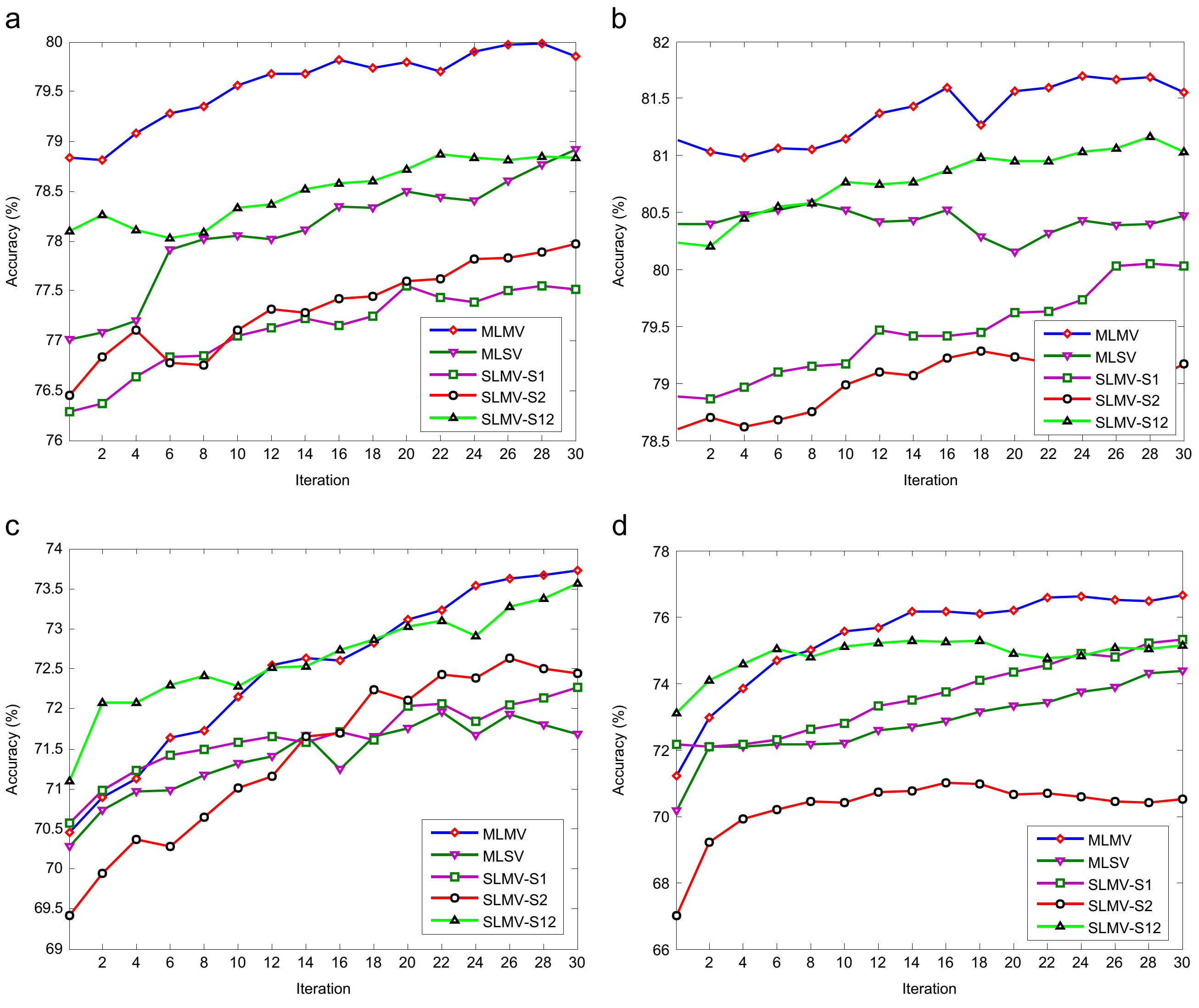


Fig. 3. Average learning curves by 3-fold cross-validation for different methods on the four different languages. (a) EnGe-Fr, (b) EnFr-Ge, (c) EnFr-Jp and (d) EnJp-Ch.

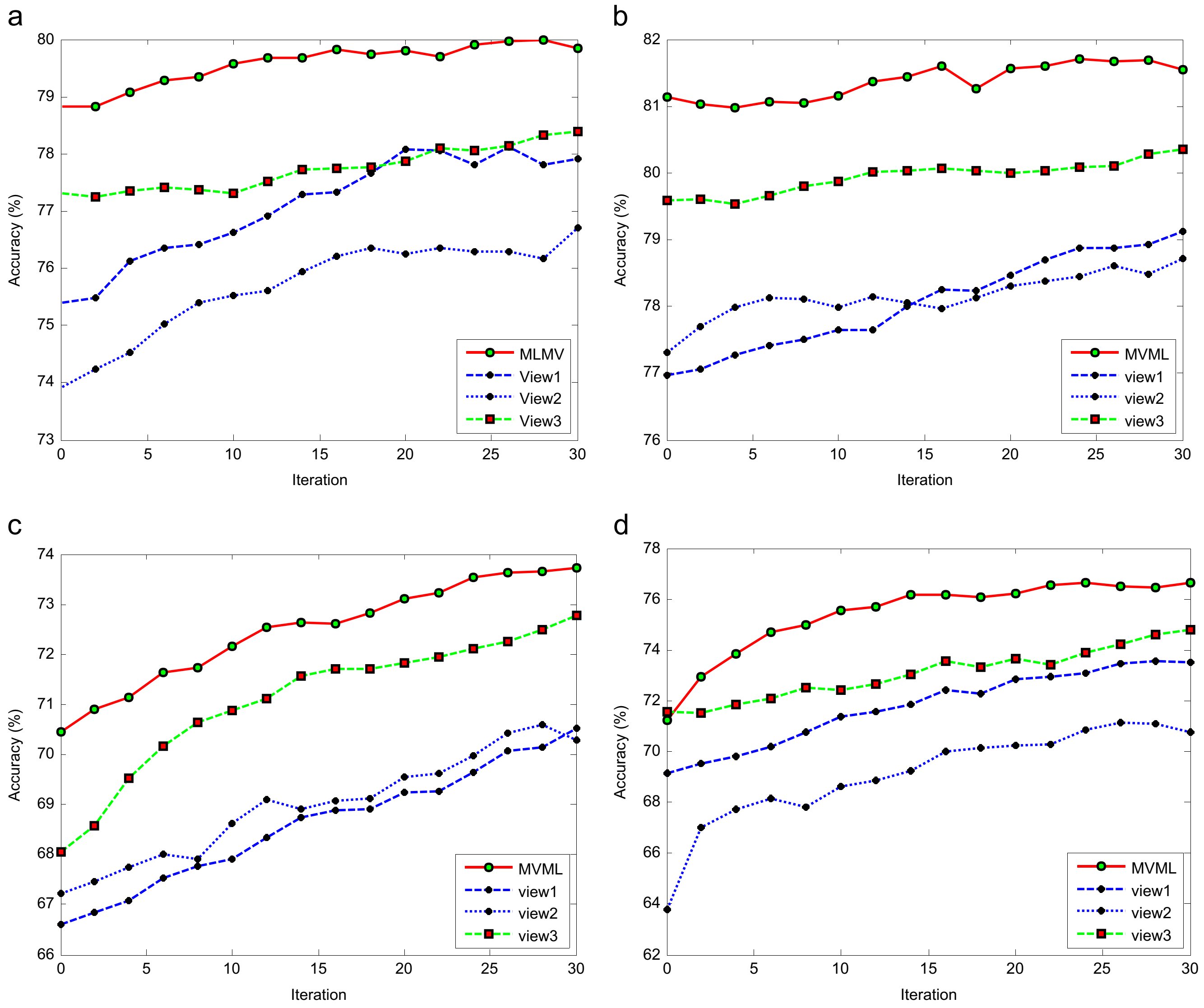


Fig. 4. Average learning curves by 3-fold cross-validation for the proposed model and each of the individual views on the four different languages. (a) EnGe-Fr, (b) EnFr-Ge, (c) EnFr-Jp and (d) EnJp-Ch.

model (e.g. SVM or Naïve Bayes) can examine the classification process from different aspects.

Acknowledgments

The Universiti Teknologi Malaysia (UTM) under research grant 03H02 and Ministry of Science, Technology & Innovations Malaysia, under research grant 4S062 are hereby acknowledged for some of the facilities utilized during the course of this research work.

References

Balahur, A., Mihalcea, R., Montoyo, A., 2014. Computational approaches to subjectivity and sentiment analysis: present and envisaged methods and applications. Comput. Speech Lang. 28, 1–6.

Balahur, A., Turchi, M., 2014. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. Comput. Speech Lang. 28, 56–75.

Banea, C., Mihalcea, R., Wiebe, J., 2010. Multilingual subjectivity: are more languages better? In: Proceedings of the 23rd International Conference on Computational Linguistics. Association for Computational Linguistics, Beijing, China, pp. 28–36.

Banea, C., Mihalcea, R., Wiebe, J., Hassan, S., 2008. Multilingual subjectivity analysis using machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Honolulu, Hawaii, pp. 127–135.

Blitzer, J., McDonald, R., Pereira, F., 2006. Domain adaptation with structural correspondence learning. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Sydney, Australia, pp. 120–128.

Blum, A., Mitchell, T., 1998. Combining labeled and unlabeled data with co-training, In: Proceedings of the Eleventh Annual Conference on Computational Learning Theory. ACM, Madison, Wisconsin, United States, pp. 92–100.

Hajmohammadi, M.S., Ibrahim, R., Selamat, A., 2014a. Bi-view semi-supervised active learning for cross-lingual sentiment classification. Inf. Process. Manag. 50, 718–732.

Hajmohammadi, M.S., Ibrahim, R., Selamat, A., 2014b. Density based active self-training for cross-lingual sentiment classification. In: Jeong, H.Y., Yen, N.Y., Park, J.J. (Eds.), Advanced in Computer Science and its Applications. Springer, Berlin Heidelberg, pp. 1053–1059.

Joachims, T., 1999. Making large-scale support vector machine learning practical, Advances in kernel methods. MIT Press, Cambridge, MA, USA, pp. 169–184.

Ku, L.W., Liang, Y.T., Chen, H.H., 2006. Opinion extraction, summarization and tracking in news and blog corpora, In: Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs.

Liu, B., 2012. Sentiment Analysis and Opinion Mining, Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, San Rafael, California, USA, pp. 1–167.

Martín-Valdivia, M.-T., Martínez-Cámara, E., Perea-Ortega, J.-M., Ureña-López, L.A., 2013. Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches. Expert Syst. Appl. 40, 3934–3942.

Mihalcea, R., Banea, C., Wiebe, J., 2007. Learning multilingual subjective language via cross-lingual projections, In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. pp. 976–983.

Moh, T.-S., Zhang, Z., 2012. Cross-lingual text classification with model translation and document translation, In: Proceedings of the 50th Annual Southeast Regional Conference. ACM, Tuscaloosa, Alabama, pp. 71–76.

Moraes, R., Valiati, J.F., Gavião Neto, W.P., 2013. Document-level sentiment classification: an empirical comparison between SVM and ANN. Expert Syst. Appl. 40, 621–633.

Pan, J., Xue, G.-R., Yu, Y., Wang, Y., 2011. Cross-lingual sentiment classification via Bi-view non-negative matrix tri-factorization. In: Huang, J., Cao, L., Srivastava, J. (Eds.), Advances in Knowledge Discovery and Data Mining. Springer, Berlin/ Heidelberg, pp. 289–300.

Pang, B., Lee, L., Vaithyanathan, S., 2002. Thumbs up? Sentiment classification using machine learning techniques, In: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pp. 79–86.

Prettenhofer, P., Stein, B., 2010. Cross-language text classification using structural correspondence learning, In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Uppsala, Sweden, pp. 1118–1127.

Shi, L., Mihalcea, R., Tian, M., 2010. Cross language text classification by model translation and semi-supervised learning, In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Cambridge, Massachusetts, pp. 1057–1067.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M., 2011. Lexicon-based methods for sentiment analysis. Comput. Linguist. 37, 267–307.

Turney, P.D., 2002. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews, In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, Philadelphia, Pennsylvania, pp. 417–424.

Wan, X., 2008. Using bilingual knowledge and ensemble techniques for unsupervised Chinese sentiment analysis, In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Honolulu, Hawaii, pp. 553–561.

Wan, X., 2009. Co-training for cross-lingual sentiment classification, 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. Association for Computational Linguistics, Suntec, Singapore, pp. 235–243.

Wan, X., 2011. Bilingual co-training for sentiment classification of Chinese product reviews. Comput. Linguist. 37, 587–616.

Xia, R., Zong, C., Li, S., 2011. Ensemble of feature sets and classification algorithms for sentiment classification. Inf. Sci. 181, 1138–1152.

Zhou, S., Chen, Q., Wang, X., 2013. Active deep learning method for semi-supervised sentiment classification. Neurocomputing 120, 536–546.

Zhou, Z.-H., Li, M., 2010. Semi-supervised learning by disagreement. Knowl. Inf. Syst. 24, 415–439.