Emoji-Powered Representation Learning for Cross-Lingual Sentiment Classification

Zhenpeng Chen
Key Lab of High-Confidence Software Technology, MoE (Peking University), Beijing, China
czp@pku.edu.cn

Sheng Shen
Key Lab of High-Confidence Software Technology, MoE (Peking University), Beijing, China
University of California, Berkeley, USA
sheng.s@berkeley.edu

Ziniu Hu
Key Lab of High-Confidence Software Technology, MoE (Peking University), Beijing, China
University of California, Los Angeles, USA
bull@cs.ucla.edu

Xuan Lu
Key Lab of High-Confidence Software Technology, MoE (Peking University), Beijing, China
luxuan@pku.edu.cn

Qiaozhu Mei
School of Information, University of Michigan, Ann Arbor, USA
qmei@umich.edu

Xuanzhe Liu*
Key Lab of High-Confidence Software Technology, MoE (Peking University), Beijing, China
xzl@pku.edu.cn

ABSTRACT

Sentiment classification typically relies on a large amount of labeled data. In practice, the availability of labels is highly imbalanced among different languages, e.g., more English texts are labeled than texts in any other languages, which creates a considerable inequality in the quality of related information services received by users speaking different languages. To tackle this problem, cross-lingual sentiment classification approaches aim to transfer knowledge learned from one language that has abundant labeled examples (i.e., the source language, usually English) to another language with fewer labels (i.e., the target language). The source and the target languages are usually bridged through off-the-shelf machine translation tools. Through such a channel, cross-language sentiment patterns can be successfully learned from English and transferred into the target languages. This approach, however, often fails to capture sentiment knowledge specific to the target language, and thus compromises the accuracy of the downstream classification task. In this paper, we employ emojis, which are widely available in many languages, as a new channel to learn both the cross-language and the language-specific sentiment patterns. We propose a novel representation learning method that uses emoji prediction as an instrument to learn respective sentiment-aware representations for each language. The learned representations are then integrated to facilitate cross-lingual sentiment classification. The proposed method demonstrates state-of-the-art performance on benchmark datasets, which is sustained even when sentiment labels are scarce.

CCS CONCEPTS

• **Information systems → Sentiment analysis**.

*Corresponding author: Xuanzhe Liu (xzl@pku.edu.cn).

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313600

KEYWORDS

Emoji; cross-lingual analysis; sentiment classification

ACM Reference Format:
Zhenpeng Chen, Sheng Shen, Ziniu Hu, Xuan Lu, Qiaozhu Mei, and Xuanzhe Liu. 2019. Emoji-Powered Representation Learning for Cross-Lingual Sentiment Classification. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3308558.3313600

1 INTRODUCTION

Sentiment analysis has become a critical topic in various research communities, including natural language processing (NLP) [22, 26], Web mining [23, 38], information retrieval [14, 57], ubiquitous computing [30, 53], and human-computer interaction [24, 59]. Due to its effectiveness in understanding user attitudes, emotions, and even latent psychological statuses from text, sentiment analysis has been widely applied to all kinds of Web content such as blogs, Tweets, user reviews, and forum discussions, and it has been a critical component in many applications such as customer review tracking [27], sales prediction [39], product ranking [41], stock market prediction [54], opinion polling [45], recommender systems [55], personalized content delivery [29], and online advertising [52].

Similar to many other text mining tasks, existing work on sentiment analysis mainly deals with English texts [22, 26, 38, 57]. Although some efforts have also been made with other languages such as Japanese [51], sentiment analysis for non-English languages is far behind. This creates a considerable inequality in the quality of the aforementioned Web services received by non-English users, especially considering that 74.6% of Internet users are non-English speakers as of 2018 [9]. The cause of this inequality is quite simple: effective sentiment analysis tools are often built upon supervised learning techniques, and *there are way more labeled examples in English than in other languages.*

A straightforward solution is to transfer the knowledge learned from a label-rich language (i.e., the source language, usually English) to another language that has fewer labels (i.e., the target language), an approach known as *cross-lingual sentiment classification* [19].

In practice, the biggest challenge of cross-lingual sentiment classification is how to fill the linguistic gap between English and the target language. Many choose to bridge the gap through standard NLP techniques, and in particular, most recent studies have been using off-the-shelf machine translation tools to generate pseudo parallel corpora and then learn bilingual representations for the downstream sentiment classification task [50, 58, 62]. More specifically, many of these methods enforce the aligned bilingual texts to share a unified embedding space, and sentiment analysis of the target language is conducted in that space.

Although this approach looks sensible and easily executable, the performance of these machine translation-based methods often falls short. Indeed, a major obstacle of cross-lingual sentiment analysis is the so-called *language discrepancy* problem [19], which machine translation does not tackle well. More specifically, sentiment expressions often differ a lot across languages. Machine translation is able to retain the general expressions of sentiments that are shared across languages (e.g., "*angry*" or "怒っている" for negative sentiment), but it usually loses or even alters the sentiments in language-specific expressions [44]. As an example, in Japanese, the common expression "湯水のように使う" indicates a negative sentiment, describing the excessive usage or waste of a resource. However, its translation in English, "*use it like hot water*," not only loses the negative sentiment but also sounds odd.

The reason behind this pitfall is easy to explain: machine translation tools are usually trained on parallel corpora that are built in the first place to capture patterns shared across languages instead of patterns specific to individual languages. In other words, the problem is due to the failure to retain language-specific sentiment knowledge when unilaterally pursuing generalization across languages. A new bridge needs to be built beyond machine translation, which not only transfers "general sentiment knowedge" from the source language but also captures "private sentiment knowledge" of the target language. *That bridge can be built with emojis.*

In this paper, we tackle the problem of cross-lingual sentiment analysis by employing *emojis* as an instrument. Emojis are considered an emerging ubiquitous language used worldwide [16, 40]; in our approach they serve both as a proxy of sentiment labels and as a bridge between languages. Their functionality of expressing emotions [21, 34] motivates us to employ emojis as complementary labels for sentiments, while their ubiquity [16, 40] makes it feasible to learn emoji-sentiment representations for almost every active language. Coupled with machine translation, the cross-language patterns of emoji usage can complement the pseudo parallel corpora and narrow the language gap, and the language-specific patterns of emoji usage help address the language discrepancy problem.

We propose ELSA, a novel framework of *Emoji-powered representation learning for cross-Lingual Sentiment Analysis*. Through ELSA, language-specific representations are first derived based on modeling how emojis are used alongside words in each language. These per-language representations are then integrated and refined to predict the rich sentiment labels in the source language, through the help of machine translation. Different from the mandatorily aligned bilingual representations in existing studies, the joint representation learned through ELSA catches not only the general sentiment patterns across languages, but also the language-specific patterns. In this way, the new representation and the downstream tasks are no longer dominated by the source language.

We evaluate the performance of ELSA on a benchmark Amazon review dataset that has been used in various cross-lingual sentiment classification studies [50, 58, 62]. The benchmark dataset covers nine tasks combined from three target languages (i.e., Japanese, French, and German) and three domains (i.e., book, DVD, and music). Results indicate that ELSA outperforms existing approaches on all of these tasks in terms of classification accuracy. Experiments also show that the emoji-powered model works well even when the volume of unlabeled and labeled data are rather limited. To evaluate the generalizability of ELSA, we also apply the method to Tweets, which again demonstrates state-of-the-art performance. In summary, the major contributions of this paper are as follows:

- To the best of our knowledge, this is the first study that leverages emojis as an instrument in cross-lingual sentiment classification. We demonstrate that emojis provide not only surrogate sentiment labels but also an effective way to address language discrepancy.
- We propose a novel representation learning method to incorporating language-specific knowledge into cross-lingual sentiment classification, which uses an attention-based Long Short-Term Memory (LSTM) model to capture sentiments from emoji usage.
- We demonstrate the effectiveness and efficiency of ELSA for cross-lingual sentiment classification using multiple large-scale datasets. ELSA significantly improves the state-of-the-art results on the benchmark datasets.[1]
- The use of emojis as a bridge provides actionable insights into other Web mining applications that suffer from similar problem of inequality among languages.

The rest of this paper is organized as follows. Section 2 presents the related work. Section 3 formulates the problem and presents the proposed approach (ELSA) to cross-lingual representation learning. Section 4 evaluates ELSA and analyzes the effectiveness of emojis in the learning process. Section 5 discusses the scalability and generalizability of ELSA, followed by concluding remarks in Section 6.

2 RELATED WORK

We start with a summary of existing literature related to our study.

Emojis. Emojis, also known as ideograms or smileys, can be used as compact expressions of objects, topics, and emotions. Being encoded in Unicode, they have no language barriers and are diffused on the Internet rapidly [40]. The prevalence of emojis has attracted researchers from various research communities such as NLP, ubiquitous computing, human-computer interaction, multimedia, and Web mining [12, 16, 20, 21, 34, 40, 43]. Many efforts have been devoted to studying their usage across platforms [43], across genders [20], across languages [16], and across cultures [40]. The various non-verbal functions of emojis play an important role in their wide adoption. Emojis are used to replace content words, express situational and additional emotions, adjust tones, express intimacy, etc. [21, 34]. In particular, expressing sentiment is demonstrated to be the most popular intention for using emojis [34], so that emojis

[1]The benchmark datasets, scripts, and pre-trained models are available at https://github.com/sInceraSs/ELSA.

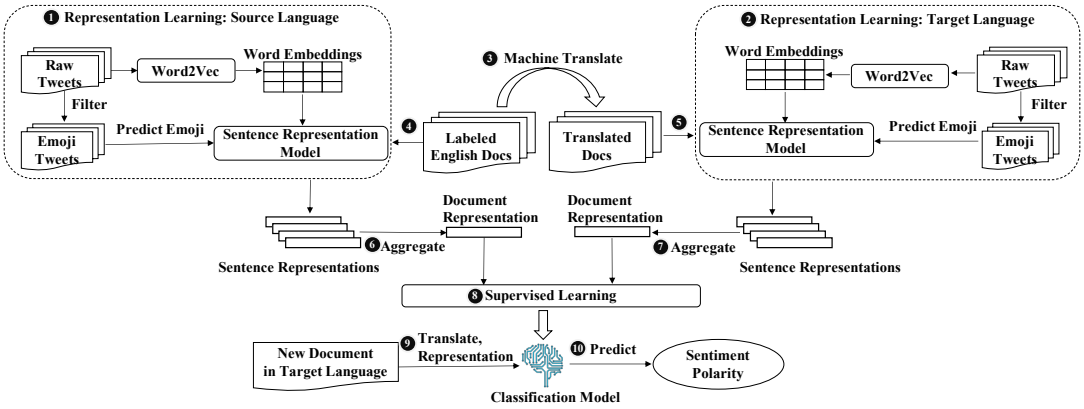


Figure 1: The workflow of ELSA.

can be used as effective proxies for sentiment polarities [26]. Considering the ubiquitous usage of emojis across languages and their functionality of expressing sentiments, we make the first effort to use emojis as an instrument to improve cross-lingual sentiment analysis.

Textual Sentiment Analysis. Sentiment analysis is a classical NLP task aiming to study the emotions, opinions, evaluations, appraisals, and attitudes of people from text data [37]. Many widely used tools, such as SentiStrength [56] and LIWC [49], simply aggregate the polarity of individual words to determine the overall sentiment score of a text. Better performance of sentiment classification is often obtained through supervised machine learning [46]. Recently, with the emergence of deep learning techniques, many researchers have attempted to use advanced neural network models for sentiment analysis [60]. Supervised machine learning methods, including deep learning models, usually require a large volume of labeled data for training. In reality, however, high-quality sentiment labels are often scarce due to the labor-consuming and error-prone human annotation process [26]. To address this limitation, researchers have used sentimental hashtags and emoticons as weak sentiment labels [22, 23]. These weak labels are usually language/community-specific. In addition, figuring out the sentiment polarities of certain hashtags or emoticons can be hard. In recent years, emoticons have been gradually replaced by increasingly popular emojis [48], and emojis have started to be explored as proxies of sentiment labels [26]. We follow the same intuition and utilize emojis as surrogate labels to learn per-language representations. Instead of attempting to directly map emojis to sentiment polarities, however, we integrate these language-specific representations and feed them through downstream tasks to predict real, high quality sentiment labels (in the source language).

Cross-Lingual Text Classification. There is a significant imbalance in the availability of labeled corpora among different languages: more in English, and much fewer in other languages. Cross-lingual learning is a common approach to tackling this problem in various text mining tasks such as Web page classification [36], topic categorization [61], and sentiment analysis [50, 58, 62]. Many researchers divide cross-lingual learning process into two stages: first encoding texts in the source and the target languages into continuous representations, and then utilizing these representations for the final classification task in the target language [50, 58, 62]. To bridge the linguistic gap between the source and the target languages, most studies introduce a translation oracle to project different languages' representations into a unified space at different (e.g., word or document) levels [18, 50, 58, 62]. The performance of these methods thus heavily depends on the quality of the machine translation tools and the pseudo parallel corpora they generate. Unfortunately, different from topical words, emotional language patterns like sentiment (or sarcasm, humor), which present strong language-specific characteristics, cannot be easily transferred in this way. We utilize the easily accessible emoji-texts to incorporate both cross-language and language-specific knowledge into the representations of the source and the target languages. The implicit sentiment knowledge encoded in the usage of diverse emojis solves both the label imbalance and the language discrepancy problems.

3 THE ELSA APPROACH

To better illustrate the workflow of ELSA, we first give a formulation of our problem. Cross-lingual sentiment classification aims to use the labeled data in a source language (i.e., English) to learn a model that can classify the sentiment of test data in a target language. In our setting, besides labeled English documents (L_S), we also have large-scale unlabeled data in English (U_S) and in the target language (U_T). Furthermore, there exist unlabeled data containing emojis, both in English (E_S) and in the target language (E_T). In practice, these unlabeled, emoji-rich data can be easily obtained from online social media such as Twitter. Our task is to build a model that can classify the sentiment polarity of document in the target language solely based on the labeled data in the source language (i.e., L_S) and the different kinds of unlabeled data (i.e., U_S, U_T, E_S and E_T). Finally, we use a held-out set of labeled documents in the target language (L_T), which can be small, to evaluate the model.

The workflow of ELSA is illustrated in Figure 1, with the following steps. In *step 1* and *step 2*, we build sentence representation models for both the source and the target languages. Specifically, for each language, we employ a large number of Tweets to learn word

embeddings (through Word2Vec [42]) in an unsupervised fashion. From these word embeddings, we learn higher-level sentence representation through predicting the emojis used in a sentence. This can be viewed as a distantly supervised learning process, where emojis serve as surrogate sentiment labels. In *step 3*, we translate each labeled English document into the target language, sentence by sentence, through *Google Translate*. Both the English sentences and their translations are fed into the representation models learned in steps 1 and 2 to obtain their per-language representations (*step 4* and *step 5*). Then in *step 6* and *step 7* we aggregate these sentence representations back to form two compact representations for each training document, one in English and the other in the target language. In *step 8*, we use the two representations as features to predict the real sentiment label of each document and obtain the final sentiment classifier. In the test phase, for a new document in the target language, we translate it into English and then follow the previous steps to obtain its representation (*step 9*), based on which we predict the sentiment label using the classifier (*step 10*).

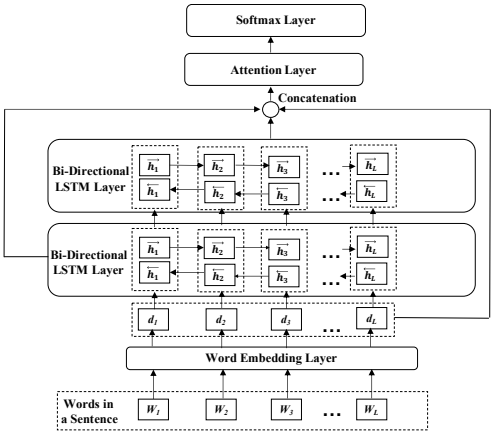


Figure 2: Network architecture for representation learning through emoji prediction.

3.1 Representation Learning

Representations of documents need to be learned before we train the sentiment classifier. Intuitively, one could simply use off-the-shelf word embedding techniques to create word representations and then average the word vectors to obtain document embeddings. Such embeddings, however, capture neither per-language nor cross-language sentiment patterns. Since emojis are widely used to express sentiments across languages, we learn sentiment-aware representations of documents using emoji prediction as an instrument. Specifically, in a distantly supervised way, we use emojis as surrogate sentiment labels and learn sentence embeddings by predicting which emojis are used in a sentence. This representation learning process is conducted separately in the source and the target languages to capture language-specific sentiment expressions.

The architecture of the representation learning model is illustrated in Figure 2. First, we pre-train low-level word embeddings using tens of millions of unlabeled Tweets (i.e., the word embedding layer). Then, we represent every single word as a unique vector and use stacked bi-directional LSTM layers and one attention layer to encode these word vectors into sentence representations. The attention layer takes the outputs of both the embedding layer and the two LSTM layers as input, through the skip-connection algorithm [31], which enables unimpeded information flow in the whole training process. Finally, the model parameters are learned by minimizing the output error of the softmax layer. The details of the architecture are elaborated below.

Word Embedding Layer. The word embeddings are pre-trained with the skip-gram algorithm [42] based on either U_S or U_T, which encode every single word into a continuous vector space. Words that commonly occur in a similar context are embedded closely in the vector space, which captures word semantic information. We leave the details of this standard Word2Vec process to the readers [42].

Bi-Directional LSTM Layer. As a special type of recurrent neural network (RNN), LSTM [33] is particularly suitable for modeling the sequential property of text data. At each step (e.g., word token), LSTM combines the current input and knowledge from the previous steps to update the states of the hidden layer. To tackle the gradient vanishing problem [32] of traditional RNNs, LSTM incorporates a gating mechanism to determine when and how the states of hidden layers can be updated. Each LSTM unit contains a memory cell and three gates (i.e., an input gate, a forget gate, and an output gate) [47]. The input and output gates control the input activations into the memory cell and the output flow of cell activations into the rest of the network, respectively. The memory cells in LSTM store the sequential states of the network, and each memory cell has a self-loop whose weight is controlled by the forget gate.

Let us denote each sentence in E_S or E_T as (x, e), where $x = [d_1, d_2, ..., d_L]$ as a sequence of *word* vectors representing the plain text (by removing emojis) and e as one emoji contained in the text. At step t, LSTM computes unit states of the network as follows:

$$
\begin{aligned}
i^{(t)} &= \sigma(U_i d_t + W_i h^{(t-1)} + b_i),\\
f^{(t)} &= \sigma(U_f d_t + W_f h^{(t-1)} + b_f),\\
o^{(t)} &= \sigma(U_o d_t + W_o h^{(t-1)} + b_o),\\
c^{(t)} &= f_t \odot c^{(t-1)} + i^{(t)} \odot tanh(U_c d_t + W_c h^{(t-1)} + b_c),\\
h^{(t)} &= o^{(t)} \odot tanh(c^{(t)}),
\end{aligned}
$$

where $i^{(t)}$, $f^{(t)}$, $o^{(t)}$, $c^{(t)}$, and $h^{(t)}$ denote the state of the input gate, forget gate, output gate, memory cell, and hidden layer at step t. W, U, b respectively denote the recurrent weights, input weights, and biases. $\odot$ is the element-wise product. We can extract the latent vector for each step t from LSTM. In order to capture the information from the context both preceding and following a word, we use the bi-directional LSTM. We concatenate the latent vectors from both directions to construct a bi-directional encoded vector h_i for every single word vector d_i, which is:

$$
\begin{aligned}
\overrightarrow{h_i} &= \overrightarrow{LSTM}(d_i), i \in [1, L],\\
\overleftarrow{h_i} &= \overleftarrow{LSTM}(d_i), i \in [L, 1],\\
h_i &= [\overrightarrow{h_i}, \overleftarrow{h_i}].
\end{aligned}
$$

Attention Layer. We employ a skip-connection that concatenates the outputs of the embedding layer and the two bi-directional LSTM

layers as the input of the attention layer. The i-th word of the input sentence can be represented as u_i:

$$u_i = [d_i, h_{i1}, h_{i2}],$$

where d_i, h_{i1}, and h_{i2} denote the encoded vectors of words extracted in the word embedding layer and the first and second bi-directional LSTMs, respectively. Since not all words contribute equally to predicting emojis or expressing sentiments, we employ the attention mechanism [13] to determine the importance of every single word. The attention score of the i-th word is calculated by

$$a_i = \frac{exp(W_a u_i)}{\sum_{j=1}^{L} exp(W_a u_j)},$$

where W_a is the weight matrix used by the attention layer. Then each sentence can be represented as the weighted sum of all words in it, using the attention scores as weights. That is,

$$v = \sum_{i=1}^{L} a_i u_i.$$

Softmax Layer. The sentence representation is then transferred into the softmax layer, which returns a probability vector Y. Each element of this vector indicates the probability that this sentence contains a specific emoji. The i-th element of the probability vector is calculated as:

$$y_i = \frac{exp(v^T w_i + b_i)}{\sum_{j=1}^{K} exp(v^T w_j + b_j)},$$

where w_i and b_i define the weight and bias of the i-th element. Finally, we learn the model parameters by minimizing the cross entropy between the output probability vectors and the one-hot vectors of the emoji contained in each sentence. After learning the parameters, we can extract the output of the attention layer to represent each input sentence. Through this emoji-prediction process, words with distinctive sentiments can be identified, and the plain text surrounding the same emojis will be represented similarly. Given the fact that the sentiment labels are limited, once the emoji-powered sentence representations are trained, they are locked in the downstream sentiment prediction task to avoid over-fitting.

3.2 Training the Sentiment Classifier

Based on the pre-trained, per-language sentence representations, we then learn document representations and conduct cross-lingual sentiment classification.

First, for each English document $D_s \in L_S$, we use the pre-trained English representation model to embed every single sentence in it. Second, we aggregate these sentence representations to derive a compact document representation. Because different parts of a document contribute differently to the overall sentiment, we once again adopt the attention mechanism here. Supposing the sentence vectors as v_i, we calculate the document vector r_s as:

$$r_s = \sum_{i=1}^{N} \beta_i v_i, \text{ where}$$
$$\beta_i = \frac{exp(W_b v_i)}{\sum_{j=1}^{N} exp(W_b v_j)},$$

Table 1: The sizes of the Tweets and emoji-Tweets.

Language	English	Japanese	French	German
Raw Tweets	39.4M	19.5M	29.2M	12.4M
Emoji-Tweets	6.6M	2.9M	4.4M	2.7M

where W_b is the weight matrix of the attention layer and β_i is the attention score of the i-th sentence in the document. Next, we use *Google Translate* to translate D_s into the target language (D_t). We then leverage the pre-trained target-language representation model to form representations for each translated document following the same process above. Supposing the text representations of D_s and D_t are r_s and r_t respectively, we concatenate them into a joint representation $r_c = [r_s, r_t]$, which contains sentiment knowledge from both English and the target language, ensuring that our model is not dominated by the labeled English documents. Finally, we input r_c into an additional softmax layer to predict the real sentiment label of D_s.

3.3 Sentiment Classification for Target Language

When we receive an unlabeled document in L_T, we first translate it into English. Based on the representation models trained above, the original document and its English translation can be represented as r_t and r_s. We represent this document as $[r_s, r_t]$ and input it into the classifier, which outputs a predicted sentiment polarity.

4 EVALUATION

In this section, we evaluate the effectiveness and efficiency of ELSA using standard benchmark datasets for cross-lingual sentiment classification as well as a large-scale corpus of Tweets.

4.1 The Dataset

The labeled data (L_S for training and L_T for testing) used in our work are from the Amazon review dataset [3] created by Prettenhofer and Stein [50]. This dataset is representative and used in a variety of cross-lingual sentiment classification work [50, 58, 62]. It covers four languages (i.e., English, Japanese, French, and German) and three domains (i.e., book, DVD, and music). For each combination of language and domain, the dataset contains 1,000 positive reviews and 1,000 negative reviews. We select English as the source language and the other three as the target languages. Therefore, we can evaluate our approach on nine tasks in total (i.e., combinations of the three domains and three target languages). For each task, we use the 2,000 labeled English reviews in the corresponding domain for training and the 2,000 labeled reviews in each target language for evaluation. The translations of the test reviews are already provided in this dataset, so we only need to translate the English reviews into target languages.

To achieve unlabeled data (U_S and U_T), we collect a sample of English, Japanese, French, and German Tweets between September 2016 and March 2018. All collected Tweets are used to train the word embeddings. As emojis are widely used on Twitter [48], we are able to extract emoji-labeled Tweets, which are used to learn emoji-powered sentence representations. For each language, we extract Tweets containing the top 64 emojis used in this language.

As many Tweets contain multiple emojis, for each Tweet, we create separate examples for each unique emoji used in it to make the emoji prediction a single-label classification task instead of more complicated multi-label classification.

We then conduct the following preprocessing procedures for the documents. We remove all Retweets, and Tweets that contain URLs, to ensure that words appear in their original contexts and that the meaning of the Tweets do not depend on external content. Then we tokenize all the texts (including reviews and Tweets) into words, convert them into lowercase, and shorten the words with redundant characters into their canonical forms (e.g., "cooooool" is converted to "cool"). As Japanese words are not separated by white spaces, we use a tokenization tool called *MeCab* [2] to segment Japanese documents. In addition, we use special tokens to replace mentions and numbers. The processed emoji-Tweets provide the E_S and E_T datasets, whose statistics are presented in Table 1.

4.2 Implementation Details

We learn the initial word embeddings using the skip-gram model with the window-size of 5 on the raw Tweets. The word vectors are then fine-tuned during the sentence representation learning phase. In the representation learning phase, to regularize our model, L2 regularization with parameter 10^{-6} is applied for embedding weights. Dropout is applied at the rate of 0.5 before the softmax layer. The hidden units of bi-directional LSTM layers are set as 1,024 (512 in each direction). We randomly split the emoji-Tweets into the training, validation, and test sets in the proportion of 7:2:1. Accordingly, we use early stopping [17] to tune hyperparameters based on the validation performance through 50 epochs, with mini-batch size of 250. We used the Adam algorithm [35] for optimization, with the two momentum parameters set to 0.9 and 0.999, respectively. The initial learning rate was set to 10^{-3}. In the phase of training the sentiment classifier, for exhaustive parameter tuning, we randomly select 90% of the labeled data as the training set and the remaining 10% as the validation set. The whole framework is implemented with TensorFlow [11].

4.3 Baselines and Accuracy Comparison

To evaluate the performance of ELSA, we employ three representative baseline methods for comparison:

MT-BOW uses the bag-of-words features to learn a linear classifier on the labeled English data [50]. It uses *Google Translate* to translate the test data into English and applies the pre-trained classifier to predict the sentiment polarity of the translated documents.

CL-RL is the word-aligned representation learning method proposed by Xiao and Guo [58]. It constructs a unified word representation that consists of both language-specific components and shared components, for the source and the target languages. To establish connections between the two languages, it leverages *Google Translate* to create a set of critical parallel word pairs, and then it forces each parallel word pair to share the same word representation. The document representation is computed by taking the average over all words in the document. Given the representation as features, it trains a linear SVM model using the labeled English data.

Table 2: The accuracy of ELSA (standard deviations in parentheses) and baseline methods on the nine benchmark tasks.

Language	Domain	MT-BOW	CL-RL	BiDRL	ELSA
Japanese	Book	0.702	0.711	0.732	**0.783 (0.003)**
	DVD	0.713	0.731	0.768	**0.791 (0.004)**
	Music	0.720	0.744	0.788	**0.808 (0.005)**
French	Book	0.808	0.783	0.844	**0.860 (0.002)**
	DVD	0.788	0.748	0.836	**0.857 (0.002)**
	Music	0.758	0.787	0.825	**0.860 (0.002)**
German	Book	0.797	0.799	0.841	**0.864 (0.001)**
	DVD	0.779	0.771	0.841	**0.861 (0.001)**
	Music	0.772	0.773	0.847	**0.878 (0.002)**

BiDRL is the document-aligned representation learning method proposed by Zhou *et al.* [62]. It uses *Google Translate* to create labeled parallel documents and forces the pseudo parallel documents to share the same embedding space. It also enforces constraints to make the document vectors associated with different sentiments fall into different positions in the embedding space. Furthermore, it forces documents with large textual differences but the same sentiment to have similar representations. After this representation learning process, it concatenates the vectors of one document in both languages and trains a logistic regression sentiment classifier.

As the benchmark datasets have quite balanced positive and negative reviews, we follow the aforementioned studies to use accuracy as an evaluation metric. All the baseline methods have been evaluated with exactly the same training and test data sets used in previous studies [62], so we make direct comparisons with their reported results. Unfortunately, we cannot obtain the individual predictions of these methods, so we are not able to report the statistical significance (such as McNemar's test [25]) of the difference between these baselines and ELSA. To alleviate this problem and get robust results, we run ELSA 10 times with different random initiations and summarize its average accuracy and standard deviation in Table 2, as well as the reported performance of the baselines.

As illustrated in Table 2, ELSA outperforms all three baseline methods on all nine tasks. Looking more closely, the performance of all methods in Japanese sentiment classification is worse than in French and German tasks. According to the language systems defined by ISO 639 [10], English, French, and German belong to the same language family (i.e., Indo-European), while Japanese belongs to the Japonic family. In other words, French and German are more in common with English, and it is expected to be easier to translate English texts into French and German and transfer the sentiment knowledge from English to them. Therefore, in fact, Japanese tasks are most difficult and none of the previous methods have been able to achieve an accuracy above 0.8. It is encouraging to find that ELSA achieves an accuracy of 0.808 on the Japanese music task and an accuracy close to 0.8 (0.791) on the Japanese DVD task. The 0.783 accuracy on the book task is also non-negligible as it improves on the best existing model by almost 7 percent. In addition, although the French and German tasks are a little easier than the Japanese ones, none of the existing approaches can achieve an accuracy over 0.85 on any of the six tasks. However, our approach can achieve a mean accuracy higher than 0.85 on all of the six tasks.

Next, we compare the results more thoroughly and further demonstrate the advantages of our approach. As is shown, the representation learning approaches (CL-RL, BiDRL, and ELSA) all

outperform the shallow method MT-BOW on most tasks. This is reasonable as representation learning approaches embed words into high-dimensional vectors in a continuous semantic space and thus overcome the feature sparsity issue of traditional bag-of-words approaches. Furthermore, we observe that the document-level representation approaches (BiDRL and ELSA) outperform the word-level CL-RL. This indicates that incorporating document-level information into representations is more effective than focusing on individual words. Finally, ELSA outperforms the BiDRL on all tasks. In order to narrow the linguistic gap, BiDRL leverages only pseudo parallel texts to learn the common sentiment patterns between languages. Besides the pseudo parallel texts, ELSA also learns from the emoji usage in both languages. On the one hand, as a ubiquitous emotional signal, emojis are adopted across languages to express common sentiment patterns, which can complement the pseudo parallel corpus. On the other hand, the language-specific patterns of emoji usage help incorporate the language-specific knowledge of sentiments into the representation learning, which can benefit the downstream sentiment classification in the target language. As a next step, we explore the role of emojis in the learning process with a more comprehensive investigation.

4.4 The Power of Emojis

To further evaluate the contribution of emojis in ELSA, we conduct subsequent experiments to investigate the effects of emojis from three perspectives, i.e., overall performance, effectiveness of representation learning, and text comprehension.

4.4.1 Overall Performance. To understand how emojis affect cross-lingual sentiment classification in general, a straightforward idea is to remove the emoji-prediction phase and compare simplified versions of ELSA:

N-ELSA removes the emoji-prediction phase of both languages and directly uses two attention layers to realize the transformation from word vectors to the final document representation. There is no emoji data used in this model.

T-ELSA removes the emoji-based representation learning on the English side. It uses the emoji-powered representations for the target language and translates labeled English documents into the target language to train a sentiment classifier for the target language. This model only leverages emoji usage in the target language.

S-ELSA removes the emoji-based representation learning in the target language. It uses the emoji-powered representations of English and trains a sentiment classifier based on labeled English documents. Documents in the target language are first translated into English and then classified. This model only leverages emoji usage in the source language (i.e., English).

Test accuracy of these models is illustrated in Table 3. We find that ELSA outperforms N-ELSA on all nine tasks. N-ELSA is only a little better than uniform guess (50%) since it learns the common patterns between languages only from pseudo parallel texts and does not incorporate sentiment information effectively. An alternative conjecture is that 2,000 reviews are insufficient to train such a complex model, which may have led to the problem of over-fitting.

To test between the two hypotheses, we mix up the labeled reviews in English and in the target language and randomly select 2,000 examples from the mixed set for training and use the remaining samples as a new test set. All other settings of the experiment are kept the same except for the new train/test split. Trained and tested in this way, the accuracy of N-ELSA becomes acceptable, with an average accuracy of 0.777 on all tasks. This indicates that over-fitting might not have been the major reason, while language discrepancy might be. Indeed, N-ELSA can still work well if we effectively incorporate cross-language sentiment information into the training process. More specifically, the original N-ELSA is dominated by English sentiment information learned from pseudo parallel texts and fails to generalize to the target language correctly. When we input the sentiment information (labeled documents) of both English and the target language into the model, performance improves. Unfortunately, in a cross-lingual sentiment classification setting, we can not acquire enough labels in the target language. Emojis help the model capture generalizable sentiment knowledge, even if there is no labeled example for training in the target language.

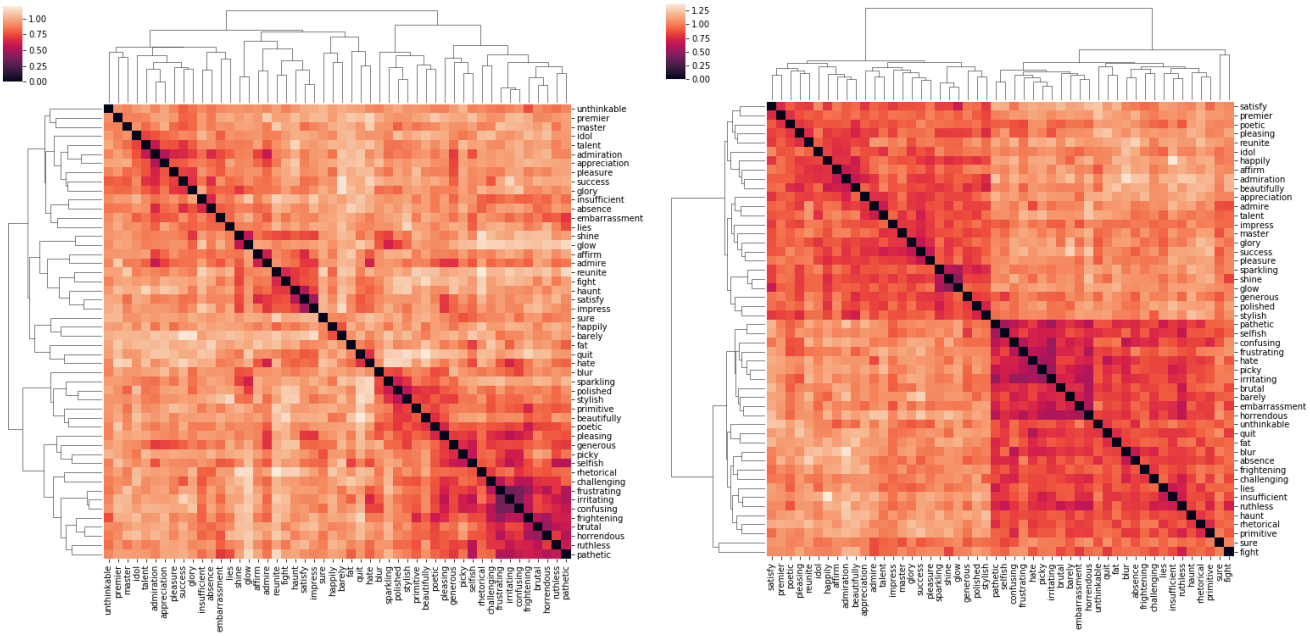
Table 3: Performance of ELSA and its simplified versions.

Language	Domain	N-ELSA	T-ELSA	S-ELSA	ELSA
Japanese	Book	0.527*	0.742*	0.753*	0.783
	DVD	0.507*	0.756*	0.766*	0.791
	Music	0.513*	0.792*	0.778*	0.808
French	Book	0.505*	0.821*	0.850*	0.860
	DVD	0.507*	0.816*	0.843*	0.857
	Music	0.503*	0.811*	0.848*	0.860
German	Book	0.513*	0.804*	0.848*	0.864
	DVD	0.521*	0.790*	0.849*	0.861
	Music	0.513*	0.818*	0.863*	0.878

* indicates the difference between ELSA and its simplified versions is statistically significant ($p < 0.05$) by McNemar's test.

In addition, ELSA also consistently achieves better accuracy compared to T-ELSA and S-ELSA on all tasks (McNemar's test [25] is performed and the differences are all statistically significant at the 5% level). The superiority of ELSA shows that only extracting sentiment information from one language is not enough for the cross-lingual sentiment task and that incorporating language-specific knowledge for both languages is critical to the model's performance. Indeed, S-ELSA fails to capture sentiment patterns in the target language; and T-ELSA falls short in extracting transferable sentiment patterns from English (indicating that emojis are still beneficial even if there are sentiment labels in a language).

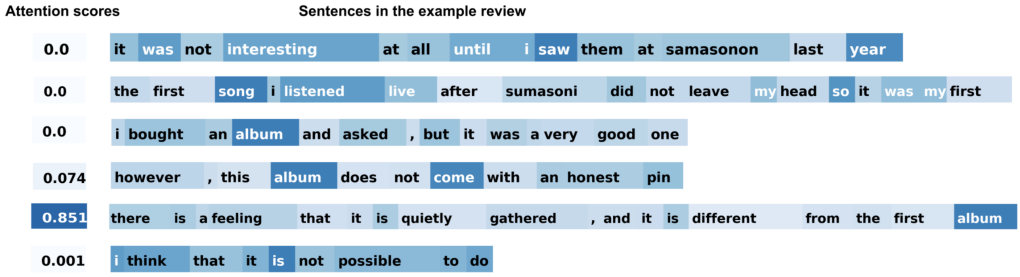
4.4.2 Effectiveness of Representation Learning. To better understand the sentiment information learned through the emoji usage, we then conduct an empirical experiment at the word representation level. Recall that after the word embedding phase, each individual word can be represented by a unique vector and that these word vectors are then fine-tuned in the emoji-prediction phase. Next, we would like to evaluate whether sentiment information is better captured by the new word representations under the effects of emojis. We sample 50 English words with distinct sentiments from the MPQA subjectivity lexicon [1] based on their frequencies in our corpus. These words are manually labeled in terms of positive or negative polarity from MPQA, and we regard these labels as the ground-truth for further evaluation.



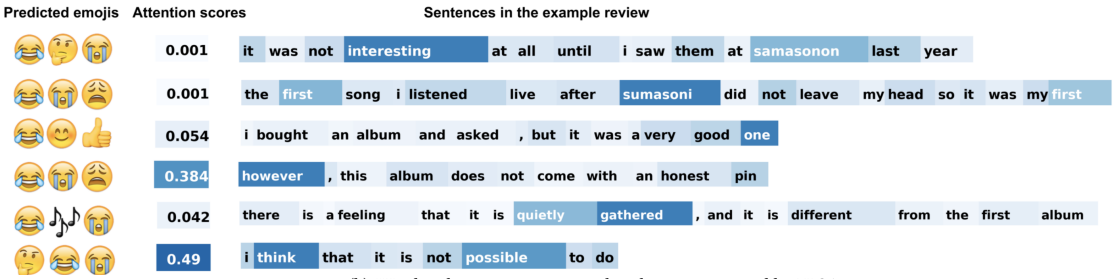
(a) Clusters of selected words with Word2Vec representations.

(b) Clusters of selected words with emoji-powered representations.

Figure 3: Comparison of word representations *with* and *without* emoji prediction.



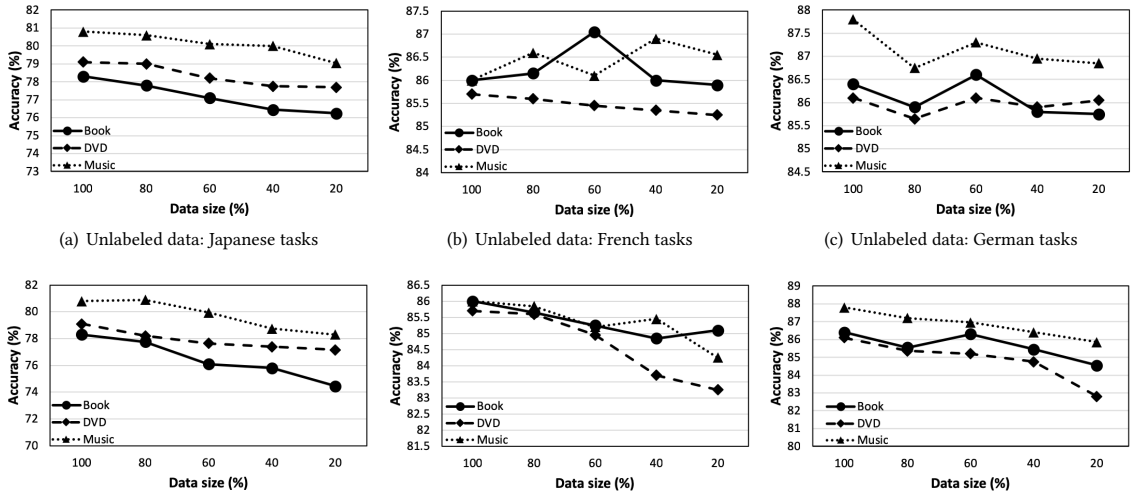
(a) Word and sentence attention distribution generated by N-ELSA.



(b) Word and sentence attention distribution generated by ELSA.

Figure 4: Case study: Effect of emojis on text comprehension.

We expect that an informative representation can embed words with same sentiment polarity closely in the vector space. To measure and illustrate the similarity, we calculate the similarity score between every two words using the cosine of the corresponding embedding vectors. Based on the cosine similarity, we perform a hierarchical clustering [15] and visualize the clustering results in Figure 3. The color scale of each cell indicates the similarity between the two words. The darker the cell, the more similar the representations of the two words.



(a) Unlabeled data: Japanese tasks (b) Unlabeled data: French tasks (c) Unlabeled data: German tasks

(d) Labeled English data: Japanese tasks (e) Labeled English data: French tasks (f) Labeled English data: German tasks

Figure 5: Accuracy of ELSA when size of unlabeled and labeled data changes.

In Figure 3(a), we use naive embeddings learned by Word2Vec (no emoji), and words with different sentiments cannot be clearly separated. Many words with different sentiments are embedded closely, for example, "generous" and "picky" in the bottom right section. This indicates that shallow word embeddings do not effectively capture the sentiment information.

In contrast, in Figure 3(b), we can easily observe two clusters after the fine-tuned emoji-prediction model. The top left corner cluster contains the positive words, while the bottom right corner contains the negative words. Only one positive word, "sure," is incorrectly clustered with negative words. By checking the contexts of this word in our corpus, we find it is usually co-used with both positive and negative words, making its polarity ambiguous. The correct clustering of nearly all the words indicates that emoji usage is an effective channel to capture sentiment knowledge, which is desirable for downstream sentiment classification.

4.4.3 Text Comprehension. We then explore how the emoji-powered representations benefit text comprehension. We select a representative case that is incorrectly classified by N-ELSA but correctly classified by ELSA. This case is selected from the Japanese test samples and we use the segment of its translated English version for illustration in Figure 4. Although the whole document expresses dissatisfaction with an album, it is not that easy to identify this intent directly from each single sentence due to the translation quality and the document's complex compositions. For example, if we consider only the third sentence without context, the author seems to express a positive attitude. However, in fact, the author expresses an obviously negative attitude in the fourth and sixth sentences.

In Figure 4, we present the attention distribution of words and sentences generated by N-ELSA and ELSA, which indicates how the two models comprehend this document, or the rationale behind their classification decisions. We use the color scale of the background to indicate the attention scores of words in each sentence. The darker the word, the more it is attended to. For each sentence, we list its attention score in this document. In Figure 4(b), we also list the top 3 emojis ELSA predicts for each sentence, which may indicate its sentiment polarity predicted by ELSA.

Let us first look at Figure 4(a), which demonstrates how N-ELSA processes the sentiment information. On the word level, N-ELSA tends to focus more on neutral words like "song" or "album" instead of sentimental words. On the sentence level, an extremely high attention is placed on the fifth sentence. However, the fifth sentence describes how the album is different from the first one and it does not express the obviously negative sentiment.

In contrast, after incorporating of emojis, ELSA is able to work with a proper logic (see Figure 4(b)). ELSA places its attention to the emotional adjectives, such as "interesting" and "not possible," and contrast conjunctions such as "however." Thus, it manages to identify the sentiment of each sentence as expected, which can be further explained by the predicted emojis on the left. Besides the most popular 😂 in our corpus, 😭 and 😩 predicted for the fourth and sixth sentence indicate the negative sentiment of the author, while 👍 and 😊 in the third sentence indicate positive sentiment. Then on the sentence level, ELSA places less attention to the positive third sentence, while centering upon the fourth and the sixth sentences. Through this comparison, we can see that emojis bring additional knowledge to the text comprehension and make the attention mechanism more effective.

5 DISCUSSION

So far, we have presented the performance of ELSA on benchmark datasets and demonstrated the power of emojis in our learning process. There are some issues that could potentially affect its effectiveness and efficiency, which call for further discussion.

5.1 Sensitivity on Data Volume

As we learn text representations from large amount of Tweets, we want to investigate whether ELSA works well with a smaller volume of data. First, we investigate the size of unlabeled data. The English representation model, once learned, can be reused by any other English-target language pair. We only need to scale down the Tweets and emoji-Tweets in the target language and observe the changes in performance on benchmarks. In details, we use 80%, 60%, 40%, and 20% of the collected Tweets to re-train the target-language representation model and keep the final supervised training fixed. We summarize the results in Figures 5(a), 5(b), and 5(c). For the Japanese tasks, when we scale down the unlabeled data, the performance gets slightly worse. Comparing the results using 20% and 100% of the Tweets, the accuracy differences in three domains are 0.021, 0.014, and 0.018, respectively. For French and German, the performance fluctuates less than 0.01. Most importantly, ELSA can outperform the existing methods on all nine tasks even with the 20% unlabeled data. This indicates that even though a target language is not as actively used on Twitter, our approach still works.

Furthermore, although there are more labeled examples in English than other languages, in general, labels are still scarce. Hence, if a model can rely on even fewer labeled English documents, it is very desirable. To test this, we scale down the labeled data by 80%, 60%, 40%, and 20%. As shown in Figures 5(d), 5(e), and 5(f), the performance of ELSA slightly declines with the decrease of labels, but even with 20% labels (i.e., 400 labeled English samples), ELSA outperforms the existing methods using all 2,000 labeled samples on almost all tasks. This shows that with the help of large-scale emoji-Tweets, the model is less dependent on sentiment labels.

5.2 Generalizability

Most previous cross-lingual sentiment studies [50, 58, 62] used the Amazon review dataset for evaluation. To compare with them, we also adopt this dataset in the main experiment of this paper. Sentiment classification in other domains such as social media is also important. Can ELSA still work well in a new domain? To evaluate the generalizability of our approach, we apply ELSA to a representative type of social media data – Tweets. As Tweets are short and informal, sentiment classification for them is considered to be a big challenge [28].

As cross-lingual studies on Tweets are very limited, we take only one recent cross-lingual method (MT-CNN) proposed by Deriu *et al.* [23] for comparison. It also relies on large-scale unlabeled Tweets and a translation tool. It first trains a sentiment classifier for English and then applies it to the translations of text documents in the target language. The training process for English Tweets contains three phases. First, it uses raw Tweets to create word embeddings just like our method. Second, it leverages ":)" and ":(" as weak labels and applies a multi-layer CNN model to adapt the word embeddings. Finally, it trains the model on labeled English Tweets. This work and our work both have coverage of French and German Tweets, so we use the two as the target languages for comparison.

As the sentiment-labeled Tweets used by [23] are released in forms of Twitter IDs and some of them are no longer available now, we cannot directly compare our model to the reported results in [23]. For fair comparison, we reproduce their method on the labeled Tweets that can still be collected. Based on the pre-trained representation models of MT-CNN [5] and ELSA, we use the same labeled English Tweets to train and validate the two classifiers and then test them on the same data (i.e., labeled French and German Tweets that can be collected). We list the sizes of the labeled English, French, and German Tweets we use in Table 4. From the distribution, a naive baseline using uniform guess would achieve an accuracy of 0.451 for French and 0.628 for German.

Table 4: The sizes of labeled Tweets collected.

Dataset	Language	Positive	Neutral	Negative
Training	English [8]	5,101	3,742	1,643
Validation	English [7]	1,038	987	365
Test	French [4]	987	1,389	718
	German [6]	1,057	4,441	1,573

Table 5: Classification accuracy on French and German Tweets.

Language	ELSA	MT-CNN	Uniform Guess
French	0.696	0.535*	0.451*
German	0.809	0.654*	0.628*

*indicates the difference between ELSA and the baseline methods is statistically significant ($p < 0.05$) by McNemar's test.

Results are summarized in Table 5. The two approaches both outperform uniform guess, and ELSA outperforms the MT-CNN by 0.161 on French task and 0.155 on German task. Although we use the same training, validation, and test set for both approaches, we are still concerned about whether the pre-trained representation models have introduced unfairness. Specifically, if we have used more unlabeled Tweets for representation learning than MT-CNN, our outstanding performance may simply attribute to the size of data. To answer this question, we refer to [23] about their data size. We find that they uses 300M raw Tweets and 60M Tweets containing ":)" and ":(" for representation learning. In contrast, we only used 81M raw Tweets and 13.7M emoji-Tweets in three languages combined. Considering that emoticons are significantly less used than emojis on Twitter [48], although they use about 4.4 times more weak-labeled Tweets, these Tweets had to be collected from much more than 4.4 times of raw Tweets than ours. It is clear ELSA outperforms MT-CNN and relies less on data size.

6 CONCLUSION

As a ubiquitous emotional signal, emojis are widely adopted across languages to express sentiments. We leverage this characteristic of emojis, both using them as surrogate sentiment labels and using emoji prediction an instrument to address the language discrepancy in cross-lingual sentiment classification. We have presented ELSA, a novel emoji-powered representation learning framework, to capture both general and language-specific sentiment knowledge in the source and the target languages for cross-lingual sentiment classification. The representations learned by ELSA capture not only sentiment knowledge that generalizes across languages, but also language-specific patterns. We evaluate ELSA with comprehensive experiments on various benchmark datasets, which outperforms the state-of-the-art cross-lingual sentiment classification methods

even when the size of labeled and unlabeled data decreases. The promising results indicate that emojis may be used as an a general instrument for text mining tasks that suffer from the scarcity of labeled examples, especially in situations where an inequality among different languages presents.

ACKNOWLEDGMENT

This work was in part supported by the National Key R&D Program of China under the grant number 2018YFB1004800 and the Beijing Municipal Science and Technology Project under the grant number Z171100005117002. Qiaozhu Mei's work was supported by the National Science Foundation under grant numbers 1633370, 1131500, and 1620319. The authors would like to thank the invaluable supports from Mr. Wei Ai at University of Michigan and Ms. Jiawei Liu at Peking University. Zhenpeng Chen and Sheng Shen made equal contributions to this work. The work of Sheng Shen and Ziniu Hu was carried out when they were undergraduate students at Peking University.

REFERENCES

[1] 2005. MPQA opinion corpus. https://mpqa.cs.pitt.edu/lexicons/subj_lexicon/. (2005). Retrieved on October 22, 2018.
[2] 2006. MeCab: Yet Another Part-of-Speech and Morphological Analyzer. http://taku910.github.io/mecab. (2006). Retrieved on October 22, 2018.
[3] 2010. Webis-CLS-10. https://www.uni-weimar.de/en/media/chairs/computer-science-department/webis/data/corpus-webis-cls-10/. (2010). Retrieved on October 22, 2018.
[4] 2015. DEFT 2015: Test Corpus. https://deft.limsi.fr/2015/corpus.fr.php?lang=en. (2015). Retrieved on April 28, 2018.
[5] 2017. Deep-mlsa. https://github.com/spinningbytes/deep-mlsa. (2017). Retrieved on October 22, 2018.
[6] 2017. SB-10k: German Sentiment Corpus. https://www.spinningbytes.com/resources/germansentiment/. (2017). Retrieved on April 28, 2018.
[7] 2017. Twitter-2015test-A. http://alt.qcri.org/semeval2017/task4/index.php?id=download-the-full-training-data-for-semeval-2017-task-4. (2017). Retrieved on April 28, 2018.
[8] 2017. Twitter-2015train-A, Twitter-2016train-A, Twitter-2016dev-A, and Twitter-2016devtest-A. http://alt.qcri.org/semeval2017/task4/index.php?id=download-the-full-training-data-for-semeval-2017-task-4. (2017). Retrieved on April 28, 2018.
[9] 2018. Internet world users by language. https://www.internetworldstats.com/stats7.htm. (2018). Retrieved on October 22, 2018.
[10] 2018. ISO 639. https://www.iso.org/iso-639-language-codes.html. (2018). Retrieved on October 22, 2018.
[11] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: a system for large-scale machine learning. In *OSDI*, Vol. 16. 265–283.
[12] Wei Ai, Xuan Lu, Xuanzhe Liu, Ning Wang, Gang Huang, and Qiaozhu Mei. 2017. Untangling emoji popularity through semantic embeddings. In *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017*. 2–11.
[13] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2014*.
[14] Georgios Balikas, Simon Moura, and Massih-Reza Amini. 2017. Multitask learning for fine-grained Twitter sentiment analysis. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2017*. 1005–1008.
[15] Ziv Bar-Joseph, David K. Gifford, and Tommi S. Jaakkola. 2001. Fast optimal leaf ordering for hierarchical clustering. In *Proceedings of the Ninth International Conference on Intelligent Systems for Molecular Biology*. 22–29.
[16] Francesco Barbieri, Germán Kruszewski, Francesco Ronzano, and Horacio Saggion. 2016. How cosmopolitan are emojis?: Exploring emojis usage and meaning over different languages with distributional semantics. In *Proceedings of the 2016 ACM Conference on Multimedia Conference, MM 2016*. 531–535.
[17] Rich Caruana, Steve Lawrence, and C. Lee Giles. 2000. Overfitting in neural nets: backpropagation, conjugate gradient, and early stopping. In *Proceedings of advances in neural information processing systems 13, NIPS 2000*. 402–408.
[18] A. P. Sarath Chandar, Stanislas Lauly, Hugo Larochelle, Mitesh M. Khapra, Balaraman Ravindran, Vikas C. Raykar, and Amrita Saha. 2014. An autoencoder approach to learning bilingual word representations. In *Advances in Neural Information Processing Systems 27, NIPS 2014*. 1853–1861.
[19] Qiang Chen, Chenliang Li, and Wenjie Li. 2017. Modeling language discrepancy for cross-lingual sentiment analysis. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017*. 117–126.
[20] Zhenpeng Chen, Xuan Lu, Wei Ai, Huoran Li, Qiaozhu Mei, and Xuanzhe Liu. 2018. Through a gender lens: learning usage patterns of emojis from large-scale Android users. In *Proceedings of the 2018 World Wide Web Conference, WWW 2018*. 763–772.
[21] Henriette Cramer, Paloma de Juan, and Joel R. Tetreault. 2016. Sender-intended functions of emojis in US messaging. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services, MobileHCI 2016*. 504–509.
[22] Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Enhanced sentiment learning using Twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics, COLING 2010*. 241–249.
[23] Jan Deriu, Aurélien Lucchi, Valeria De Luca, Aliaksei Severyn, Simon Müller, Mark Cieliebak, Thomas Hofmann, and Martin Jaggi. 2017. Leveraging large amounts of weakly supervised data for multi-language sentiment classification. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017*. 1045–1052.
[24] Nicholas Diakopoulos and David A. Shamma. 2010. Characterizing debate performance via aggregated Twitter sentiment. In *Proceedings of the 28th International Conference on Human Factors in Computing Systems, CHI 2010*. 1195–1198.
[25] Thomas G Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation* 10, 7 (1998), 1895–1923.
[26] Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017*. 1615–1625.
[27] Michael Gamon. 2004. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings of 20th International Conference on Computational Linguistics, COLING 2004*.
[28] Anastasia Giachanou and Fabio Crestani. 2016. Like it or not: a survey of Twitter sentiment analysis methods. *Comput. Surveys* 49, 2 (2016), 28:1–28:41.
[29] Ryosuke Harakawa, Daichi Takehara, Takahiro Ogawa, and Miki Haseyama. 2018. Sentiment-aware personalized Tweet recommendation through multimodal FFM. *Multimedia Tools Appl.* 77, 14 (2018), 18741–18759.
[30] Kiraz Candan Herdem. 2012. Reactions: Twitter based mobile application for awareness of friends' emotions. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing, UbiComp 2012*. 796–797.
[31] M. Hermans and B. Schrauwen. 2013. Training and analysing deep recurrent neural networks. *Proceedings of advances in Neural Information Processing Systems, NIPS 2013*, 190–198.
[32] Sepp Hochreiter. 1998. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 6, 2 (1998), 107–116.
[33] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.
[34] Tianran Hu, Han Guo, Hao Sun, Thuy-vy Thi Nguyen, and Jiebo Luo. 2017. Spice up your chat: the intentions and sentiment effects of using emojis. In *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017*. 102–111.
[35] Diederik P. Kingma and Jimmy Ba. 2014. Adam: a method for stochastic optimization. *CoRR* abs/1412.6980 (2014).
[36] Xiao Ling, Gui-Rong Xue, Wenyuan Dai, Yun Jiang, Qiang Yang, and Yong Yu. 2008. Can Chinese Web pages be classified with English data source?. In *Proceedings of the 17th International Conference on World Wide Web, WWW 2008*. 969–978.
[37] Bing Liu. 2012. *Sentiment analysis and opinion mining*. Morgan & Claypool Publishers.
[38] Qiao Liu, Haibin Zhang, Yifu Zeng, Ziqi Huang, and Zufeng Wu. 2018. Content attention model for aspect based sentiment analysis. In *Proceedings of the 2018 World Wide Web Conference, WWW 2018*. 1023–1032.
[39] Yang Liu, Xiangji Huang, Aijun An, and Xiaohui Yu. 2007. ARSA: a sentiment-aware model for predicting sales performance using blogs. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2007*. 607–614.
[40] Xuan Lu, Wei Ai, Xuanzhe Liu, Qian Li, Ning Wang, Gang Huang, and Qiaozhu Mei. 2016. Learning from the ubiquitous language: an empirical analysis of emoji usage of smartphone users. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp 2016*. 770–780.
[41] Mary McGlohon, Natalie S. Glance, and Zach Reiter. 2010. Star quality: aggregating reviews to rank products and merchants. In *Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010*.
[42] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Computer Science* (2013).

[43] Hannah Jean Miller, Jacob Thebault-Spieker, Shuo Chang, Isaac L. Johnson, Loren G. Terveen, and Brent J. Hecht. 2016. "Blissfully happy" or "ready to fight": varying interpretations of emoji. In *Proceedings of the Tenth International Conference on Web and Social Media, ICWSM 2016*. 259–268.

[44] Saif M. Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. 2016. How translation alters sentiment. *J. Artif. Intell. Res.* 55 (2016), 95–130.

[45] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From Tweets to polls: linking text sentiment to public opinion time series. In *Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010*.

[46] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, EMNLP 2002*. 79–86.

[47] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013*. 1310–1318.

[48] Umashanthi Pavalanathan and Jacob Eisenstein. 2015. Emoticons vs. emojis on Twitter: a causal inference approach. *CoRR* abs/1510.08480 (2015).

[49] J. W. Pennebaker, L. E. Francis, and R. J. Booth. 1999. Linguistic inquiry and word count: LIWC. *Lawrence Erlbaum Associates Mahwah Nj* (1999).

[50] Peter Prettenhofer and Benno Stein. 2010. Cross-language text classification using structural correspondence learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL 2010*. 1118–1127.

[51] Michal Ptaszynski, Rafal Rzepka, Kenji Araki, and Yoshio Momouchi. 2012. Automatically annotating a five-billion-word corpus of Japanese blogs for affect and sentiment analysis. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis, WASSA@ACL 2012*. 89–98.

[52] Guang Qiu, Xiaofei He, Feng Zhang, Yuan Shi, Jiajun Bu, and Chun Chen. 2010. DASA: dissatisfaction-oriented advertising based on sentiment analysis. *Expert Systems with Applications* 37, 9 (2010), 6182–6191.

[53] Koustuv Saha, Larry Chan, Kaya de Barbaro, Gregory D. Abowd, and Munmun De Choudhury. 2017. Inferring mood instability on social media by leveraging ecological momentary assessments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, IMWUT* 1, 3 (2017), 95:1–95:27.

[54] Jianfeng Si, Arjun Mukherjee, Bing Liu, Qing Li, Huayi Li, and Xiaotie Deng. 2013. Exploiting topic based Twitter sentiment for stock prediction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013*. 24–29.

[55] Lihua Sun, Junpeng Guo, and Yanlin Zhu. 2019. Applying uncertainty theory into the restaurant recommender system based on sentiment analysis of online Chinese reviews. *World Wide Web* 22, 1 (2019), 83–100.

[56] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment in short strength detection informal text. *JASIST* 61, 12 (2010), 2544–2558.

[57] Fangzhao Wu, Jia Zhang, Zhigang Yuan, Sixing Wu, Yongfeng Huang, and Jun Yan. 2017. Sentence-level sentiment classification with weak supervision. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2017*. 973–976.

[58] Min Xiao and Yuhong Guo. 2013. Semi-supervised representation learning for cross-lingual text classification. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013*. 1465–1475.

[59] Jiang Yang, Lada A. Adamic, Mark S. Ackerman, Zhen Wen, and Ching-Yung Lin. 2012. The way I talk to you: sentiment expression in an organizational context. In *Proceedings of the 2012 International Conference on Human Factors in Computing Systems, CHI 2012*.

[60] Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: a survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* 8, 4 (2018).

[61] Joey Tianyi Zhou, Sinno Jialin Pan, Ivor W. Tsang, and Shen-Shyang Ho. 2016. Transfer learning for cross-language text categorization through active correspondences construction. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI 2016*. 2400–2406.

[62] Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2016. Cross-lingual sentiment classification with bilingual document representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016*. 1403–1412.