SenticVec: Toward Robust and Human-Centric Neurosymbolic Sentiment Analysis

Xulang Zhang, Rui Mao and **Erik Cambria**
College of Computing and Data Science,
Nanyang Technological University, Singapore
{xulang.zhang,rui.mao,cambria}@ntu.edu.sg

Abstract

Recent developments in natural language processing were enabled by deep neural networks, which excel in various tasks through strong data fitting and latent feature modeling abilities. However, certain challenges linked to deep nets and supervised deep learning deserve considerations, e.g., extensive computing resources, knowledge forgetting, etc. Previous research attempted to tackle these challenges individually through irrelative techniques. However, they do not instigate fundamental shifts in the learning paradigm. In this work, we propose a novel neurosymbolic method for sentiment analysis to tackle these issues. We also propose a novel sentiment-pragmatic knowledge base that places emphasis on human subjectivity within varying domain annotations. We conducted extensive experiments to show that our neurosymbolic framework for sentiment analysis stands out for its lightweight nature, robustness across domains and languages, efficient few-shot training, and rapid convergence.

1 Introduction

Deep neural networks have demonstrated remarkable capabilities in the natural language processing (NLP) domain because of their strong capacity to capture latent features and fit data. Currently, the advancement of large language models (LLMs) shows that when the neural networks become very large and deep, sufficient pre-training grants them impressive capability, making them a staple as foundation models for a variety of tasks (Bommasani et al., 2021). Despite the fascination the academic community holds for LLMs (Mao et al., 2024a), a pressing question remains: is the relentless pursuit of model enlargement truly the paramount focus for NLP? As models become larger and deeper, many machine learning problems arise. For instance, larger models normally require more computing resources to train, resulting in more carbon emission (Patterson et al., 2021).

Traditional supervised learning, one of the most significant deep learning paradigms, suffers from issues including knowledge forgetting (Kemker et al., 2018), extensive data annotation efforts (Aggarwal et al., 2018), weak cross-domain (Singhal et al., 2023) and cross-lingual (Wang et al., 2021a) adaption, and data-centric learning pipelines (Singh, 2023). For NLP, the need for deep neural networks is usually due to the complexity of modeling language, i.e., syntax (Zhang et al., 2023a) and semantics (Mao et al., 2023a) exhibiting different patterns in different contexts.

As such, the learning of complex word sequence patterns is achieved by making the model deeper and adding different learning mechanisms (Hochreiter and Schmidhuber, 1997; Vaswani et al., 2017). Furthermore, models trained with standard annotation sets where labels were aggregated based on the majority opinion of annotators could potentially disregard nuanced individual variations (Zhu et al., 2024). In subjective tasks such as affective computing, a uniform decision-making mechanism may struggle to adequately cater to individual subjective experiences. The judgment of sentiment and emotional states from different people can vary in different contexts or domains, necessitating human-centric systems.

In this research, we investigate whether the coupling of syntactic and pragmatic features can alleviate the need for semantic features, thereby significantly reducing the number of model parameters needed to retain semantic knowledge (Yeo et al., 2024). Hence, such an approach potentially tackles the challenges prevalent in deep neural networks, including heavy reliance on extensive downstream training data and computing resources (Anas et al., 2024), slow training rates, limited adaptability across domains and languages, knowledge forgetting, and the ignorance of humans' subjectivity.

Findings of the Association for Computational Linguistics ACL 2024, pages 4851–4863

Previous sentiment analysis works target to address one of these issues with different techniques, e.g., few-shot learning for eliminating the hungry of label data (He et al., 2022b); pruning neural networks or using lightweight encoders to reduce the number of learning parameters and computing resources (Ni et al., 2023); using pre-training (Devlin et al., 2019; Mao et al., 2024b) to improve the generalization of a model in different domains and languages; leveraging continues learning techniques to mitigate the knowledge forgetting (He et al., 2022a). However, these approaches do not instigate fundamental shifts in the deep learning paradigm, nor do they comprehensively resolve all the previously mentioned issues.

To alleviate the reliance on large neural networks and downstream training data while improving model performance in cross-domain and cross-language tasks, we propose a neurosymbolic method and a knowledge base for sentiment analysis[1]. Our hypothesis is that in affective computing, models can reduce the need for semantic learning by learning pragmatic knowledge and syntactic patterns, thereby reducing the number of learning parameters. and mitigating issues associated with deep neural networks.

The advantage is threefold: **a)** The exclusion of semantic learning substantially reduces the parameters of neural networks as it obviates the necessity for the model to grasp the meaning of an extensive vocabulary across diverse contexts. **b)** Incorporating multi-dimensional symbolic representations (sentiment-pragmatic knowledge) from our knowledge base enables the construction of a human-centric system that circumvents the knowledge-forgetting issue, where different dimensions signify subjective judgments from different contributors, offering the flexibility to augment the symbolic representations with new dimensions acquired from new data without changing the existing elements. **c)** By leveraging the combination of syntactic and pragmatic features as input, it is possible to alleviate issues related to learning and inference across diverse domains and languages, primarily stemming from semantic variations in different linguistic environments. This aligns with our plans to enhance current sub-symbolic AI approaches through task decomposition, panalogy, symbol grounding, and more (Cambria et al., 2023).

Our neurosymbolic framework for sentiment analysis consists of two components, a knowledge base and a neurosymbolic model. The former, termed SenticVec, consists of sentiment-pragmatic representations for 388,158 lexical units, i.e., tokens. Each representation is in vector form, where each element is a sentiment score ranging from -1 (negative) to 1 (positive), pre-trained from a public sentiment dataset. The sentiment score is automatically generated via an explainable encoder (Han et al., 2022) that signifies the contribution of each lexical unit to the sentiment prediction.

Lexical units that contribute more will receive more extreme scores. The sentiment vector of a lexical unit consists of multiple sentiment scores, learned and computed using multiple labeled sentiment datasets. The intuition is that different annotators, annotation task instructions, and text domains can yield different annotation outcomes. For a human-centric system with cross-domain robustness, it is essential to acknowledge these sentimental variations; therefore, we opt for multi-dimensional representations, integrating subjectivity instead of relying on conventional unified sentiment scores to depict sentiment-pragmatic features.

Our neurosymbolic model consists of a 3-layer multi-layer perceptron (MLP) to learn the pragmatic representations and PoS patterns. We use an existing PoS tagger to generate PoS tag sequences from the inputs. The motivation is that 1) the size of PoS tags is smaller than the vocabulary size of words; 2) different languages can share the same PoS tag set; 3) SenticVec provides useful sentiment-pragmatic representations and can be easily mapped to other languages using bilingual dictionaries. Thus, we can use a shallow neural network by leveraging symbolic representations (e.g., PoS tags and SenticVec) to process multilingual sentiment analysis tasks. This is entirely different from previous learning paradigms that were proposed for addressing the aforementioned deep learning issues. We performed thorough experiments to demonstrate the superiority of our approach in conventional supervised learning, cross-domain inference, and multilingual learning and inference for sentiment analysis. Our method achieves comparable performance to RoBERTa-base (Liu et al., 2019) while requiring 1/7723 of its training parameters, and enables faster training in the conventional supervised learning evaluation.

[1] https://github.com/senticnet/senticvec

In cross-domain inference evaluation, the SenticVec neurosymbolic model demonstrates overall better performance than RoBERTa. In the few-shot multilingual learning evaluation, SenticVec surpasses multilingual Pre-trained Language Models (PLMs) on three languages. In cross-lingual inference evaluation, SenticVec obtains comparable performance to PLMs. Furthermore, incorporating SenticVec knowledge into multilingual PLM consistently brings accuracy gains in all settings.

The contributions of this work include: (1) We develop a novel knowledge base, incorporating sentiment-pragmatic representations that prioritize human subjectivity in annotating diverse corpora from varying domains while ensuring retention of original knowledge with the influx of new labeled data. (2) Our novel neurosymbolic framework for sentiment analysis stands out for its lightweight nature, robustness across domains and languages, efficient few-shot training, and rapid convergence.

2 Related Work

Deep-learning systems have almost become the fixture of NLP tasks. Although they achieved impressive accuracy in a variety of applications, they are not without their vices. Randomly-initialized neural networks suffer from knowledge forgetting (Kemker et al., 2018; Ramasesh et al., 2021). To resolve this, pre-training is employed to help with their ability for generalizing and transferring knowledge. However, prominent PLMs such as RoBERTa and GPT (Radford et al., 2019; Brown et al., 2020) consist of a large number of parameters trained with enormous data, which consumes substantial computational resources (Bender et al., 2021), and leads to concerns about carbon emissions and environmental impact (Patterson et al., 2021; Bannour et al., 2021). On the other hand, lightweight zero-shot and few-shot learning models still face limitations in cross-domain and cross-lingual processing, e.g., negative migration, overfitting, and underfitting (Yang, 2021; Nozza, 2021).

Neurosymbolic systems for sentiment analysis leverage both subsymbolic and symbolic approaches, e.g., coupling neural networks and knowledge graphs (Cambria et al., 2024; Kocoń et al., 2022; Zhang et al., 2023b). However, existing sentiment knowledge bases (Wiebe et al., 2005; Strapparava and Valitutti, 2004; Baccianella et al., 2010; Cambria et al., 2024) assign a singular score to a lexical unit, disregarding the inherently differing perceptions people have to subjective contents (Pavlick and Kwiatkowski, 2019; Beck et al., 2020; Troiano et al., 2021), hence failing to capture the human uncertainty and subjectivity. Additionally, these knowledge bases are not domain-specific and mostly monolingual, hence less adaptable in cross-domain and multilingual settings (Hung and Chen, 2016). Furthermore, existing neurosymbolic systems mostly take a knowledge-based approach (Du et al., 2023), thus still dependent on heavy neural networks. The PLMs are the backbone, while the symbolic representations provide complementary knowledge, e.g., sense polarity (Baccianella et al., 2010), facts (Wang et al., 2021b), and commonsense (Cambria et al., 2024) that are not directly learnable from datasets by neural networks. As a result, these neurosymbolic methods did not fundamentally address some of the issues that deep-learning-based methods commonly encounter.

3 Methodology

3.1 SenticVec Knowledge Base Development

In this section, we will describe how SenticVec is constructed. First, we construct a sentiment classifier with explainable attention for determining the importance of each token to the model's sentiment prediction. The intuition is that, the more influence a token has on the sentiment inference of the input, the stronger its sentiment intensity is. As a result, the sentiment score of a token is determined by the ground truth sentiment of the input sentences it appears in, and its attention weights in these sentences. Such sentiment scores can be categorized as pragmatic knowledge, because they embody the likelihood of a lexical unit influencing the overall sentiment of a sentence.

Unlike syntactic and semantic knowledge, which pertain to structural characteristics and meanings of words respectively, this pragmatic aspect of sentiment scores demonstrates their relevance in understanding how specific words contribute to the overall sentiment conveyed in a given context. Second, we train the classifier on 19 corpora from different domains separately to acquire 19 sentiment scores for each token to form both general and PoS-tagged vector representations. We apply an averaging technique to fill in sentiment scores when a token is not present in one of the corpora. Additionally, we translate the English SenticVec lexicon into other languages using off-the-shelf machine translation systems to expand SenticVec for multilingual processing. The resulting SenticVec knowledge base contains entries for 388,158 lexical units.

3.1.1 Sentiment Score Computation

To obtain the sentiment score of a word from a given corpus, we need to measure how salient it is in terms of sentiment for the sentences it appears in. Hence, we design a simple attention-based sentiment classifier consisting of an embedding layer, an explainable attention module, and two Feed Forward Network (FFN) layers for output. The model is trained with a given corpus so that the attention module learns to assign each token with a weight that reflects its salience to sentence sentiment.

To align with our objective of constructing a lexicon, we adopt GloVe embeddings (Pennington et al., 2014) for word representations. This is motivated by the fact that in a static embedding method, the representation of a word is the same for all of its occurrences in a corpus. We denote the GloVe embedding matrix of a given lemmatized input sequence $w = (w_1, w_2, \ldots, w_L)$ as $G = (g_1, g_2, \ldots, g_L)$.

To acquire attention weights that effectively indicate word salience (Serrano and Smith, 2019), we employ an explainable attention module called Hierarchical Attention Network (HAN, Han et al., 2022), which encodes hidden states with multiple non-linear projections and ranks the most influential tokens based on attention weights. The attention module consists of I HAN layers:

$$q_i, a_i = HAN_i(G), \tag{1}$$

where vector q_i and a_i are the yielded hidden state and attention weights at the i-th HAN layer.

The query q_I produced by the last layer HAN_I is then fed into two consecutive FFN layers to produce the sentiment prediction $\hat{y}$:

$$h = ReLU(FFN_1^c(q_I)), \tag{2}$$

$$\hat{y} = Softmax(FFN_2^c(h)). \tag{3}$$

The model parameters are optimized via cross-entropy loss

$$L = CrossEntropy(\hat{y}, \tilde{y}). \tag{4}$$

After sufficient training, the sentiment score s_l of a token w_l in the context of the input sentence w is computed as the product of its min-max normalized attention weight produced by the last attention layer HAN_I of the trained classifier, and the ground-truth sentiment label $\tilde{y}$ of w:

$$s_l = \frac{a_{I,l} - min(a_I)}{max(a_I) - min(a_I)}\tilde{y}, \tag{5}$$

so that the score s_l falls within the range of $[-1, 1]$.

3.1.2 Sentiment Vector Construction

To construct vector representations that reflect the varying sentiment intensities of words in different domains from different annotators, we train the classifier in the previous section on $N \in \{1, 2, \ldots, n, \ldots, 19\}$ corpora, separately. Our selection of corpora covers different domains including Twitter, product reviews, movie reviews, finance, news, etc., with several corpora for each domain to accommodate human's innate disagreement and uncertainty in sentiment annotation. The full list of the corpora used can be found in Table 7 in Appendix C. On each corpus, the classifier is trained on the training set until the accuracy of the validation set no longer improves. To avoid the influence of negation on the connection between sentiment and word salience, we prune out all samples that contain *not* and *never*.

Given a token appears in the training set of the n-th corpus for T times, its general sentiment score s_n^g without PoS distinction is computed as

$$s_n^g = \frac{s_{n,1} + s_{n,2} + \cdots + s_{n,T}}{T}, \tag{6}$$

where the sentiment score $s_{n,t}$ in the t-th sentence is computed via Equation 5. As a result, the general sentiment vector representation s^g of the word can be assembled as

$$s^g = (s_1^g, s_2^g, \ldots, s_N^g). \tag{7}$$

We manually set the representations of the above-mentioned negation words to vectors consisting of -1s. Alternatively, if the token is not present in the n-th corpus, we assign a placeholder value $N.A.$ to s_n^g. If the token appears M times in the training set of the n-th corpus with the PoS tag p, its sentiment score s_n^p under the PoS p is calculated similarly to Equation 6 by averaging its sentiment scores in all M instances. Its PoS-p-annotated sentiment score s^p is constructed using the same method as Equation 7. A sample of SenticVec entries associated with a token is illustrated in Figure 1.

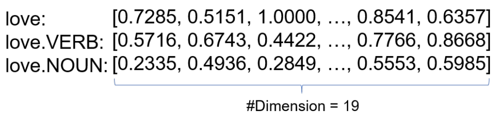

Figure 1: A sample SenticVec entries of the word *love*. We truncate the float numbers for demonstration.

After constructing vectors for all tokens in the lexicon, we fill in the $N.A.$ values using an averaging technique. For entries without PoS tags, assume s^c is the SenticVec vector of the token c, and $s^c_n = N.A.$. We find an anchor token b in the lexicon that satisfies three conditions: a) the cosine similarity of their GloVec embeddings is the largest, b) the n-th value of the anchor token's vector s^b_n is not $N.A.$, and c) in at least one dimension, the elements in both s^c and s^b are not $N.A.$. Subsequently, we fill in the value of s^c_n as

$$s^c_n = s^b_n - \frac{(s^b_{d_1} - s^c_{d_1}) + \cdots + (s^b_{d_J} - s^c_{d_J})}{J}. \quad (8)$$

J is the number of dimensions, where the values in both vectors are not $N.A.$, and $\{d_1, d_2, \ldots, d_J\}$ is the set of indices of these non-$N.A.$ dimensions. The motivation for Equation 8 is to consider the sentiment shift from different corpora. We limit the value of s^c_n to the range of [-1,1]. A simplified example is shown in Figure 2. For entries with PoS tags, we follow the same procedure except the chosen anchor entry must also have the same PoS tag. We only update the lexicon after all $N.A.$ values are filled in.

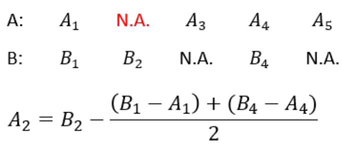

Figure 2: Simplified example of filling in N.A. value, where each vector has 5 dimensions. A_i denotes a non-N.A. value in the i-th position in A. The same goes for B_i. Red denotes the value to be filled, i.e., A_2.

3.1.3 Multilingual Lexicon Acquisition

To acquire a multilingual lexicon, we first employ Open Multilingual WordNet (OMW) (Bond and Foster, 2013) to map the lexicon into Spanish, French, and Italian, using PoS annotations as constraints when applicable. We choose the highest ranking lemma from the top synset. For tokens that are not present in OMW, we use off-the-shelf machine translation systems word2word (Choe et al., 2020) and Googletrans[2] for translation. In the case where more than one tokens are translated into the same token in the target language, we average the vectors of the original tokens as the sentiment vector of the token in the target language. When translating a PoS-annotated entry, we translate the token and keep the PoS tag as it is.

3.2 SenticVec-Based Neurosymbolic Model

Utilizing the constructed SenticVec knowledge base, we propose a lightweight, neurosymbolic model that relies on the PoS tag sequence of the input sentence for syntactic features, and the corresponding SenticVec representations for pragmatic features, as illustrated in Figure 3.

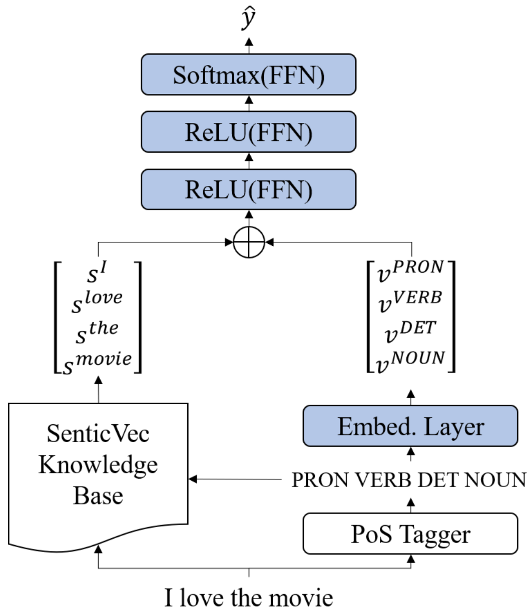

Figure 3: Our SenticVec-based neurosymbolic model. Blue denotes layers with learnable parameters.

First, given an L-long sentence, we employ an off-the-shelf PoS tagger to obtain the PoS sequence $p = (p_1, \ldots, p_L)$. An embedding layer is used to obtain PoS representations:

$$V = Embed(p). \quad (9)$$

The sentence's general SenticVec representation S^g and PoS-annotated representation S^p are constructed by finding the corresponding SenticVec entries to each lemmatized token, which are concatenated to form S. If a token has no corresponding entry, it is represented by a vector of zeros.

[2] https://github.com/ssut/py-googletrans

We then concatenate ($\oplus$) V with SenticVec representation S, and pass it to a 3-layer MLP.

$$H = ReLu(FFN^o(V \oplus S))_{\times 2}, \tag{10}$$
$$\hat{y} = Softmax(FFN_3^o(avg(H))), \tag{11}$$

where avg denotes averaging along the sequence length dimension. The model is optimized via cross-entropy loss as in Equation 4.

4 Experiments

4.1 Datasets

The Sentiment140 dataset (Go et al., 2009) (Senti.140) consists of 1.6M tweets and their corresponding sentiment labels (positive or negative). We split 10% of the original training set as the validation set. We removed the neutral samples in the original test set, since the training set does not include neutral samples. **The Review dataset** originates from the Generic Sentiment dataset[3], covering reviews from multiple domains. To align with the other datasets for cross-domain experiment setting, we remove all samples labeled as neutral from this dataset. **The Microblog dataset** (Cortis et al., 2017) originates from SemEval-2017 Task 5, where each microblog is given a polarity score ranging between -1 and 1. We repurpose it by labeling the samples with scores below 0 as negative, the rest as positive. Since the released test set does not provide ground truth labels, we split 20% of the original training set as our test set. We use the original validation set. **The Multilingual Tweet dataset** (Barbieri et al., 2021) contains 3,033 tweets per language, each labeled as positive, negative, and neutral. For a few-shot learning setting, we re-split each language subset into 5%/5%/90% for training, validation, and testing.

4.2 Baselines

We include the following baselines. **Non-pre-trained** is a 3-layer MLP model using word embeddings trained from scratch. The embedding size is the same as SenticVec. **GloVe** is a 3-layer MLP model using GloVe embeddings. **RoBERTaEmb** is a 3-layer MLP model using RoBERTa-base embeddings. **RoBERTa** consists of a RoBERTa-base encoder and a FFN output layer. **XLM-RoBERTa** (Conneau et al., 2019) consists of an XLM-RoBERTa-base encoder and a FFN output layer. The XLM-RoBERTa is a multilingual PLM. **XLM-Twitter** is the same as XLM-RoBERTa but the encoder is XLM-Twitter-base (Barbieri et al., 2021) pretrained with a large multilingual Twitter corpus. To show that SenticVec can also be effective as a complementary knowledge base, we further propose the following knowledge-based systems. **RoBERTa+SenticVec** concatenates the pooler outputs of the RoBERTa-base encoder with SenticVec representation (S) averaged along the sequence dimension, and feeds them into a FFN for output. **XLM-RoBERTa+SenticVec** use the same architecture as RoBERTa+SenticVec, except the encoder is XLM-RoBERTa-base.

[3] https://kaggle.com/datasets/akgeni/generic-sentiment-multidomain-sentiment-dataset

Dataset	Split	# Samples	# Pos	# Neu	# Neg
Senti	Trn.	1,440,144	720,098	0	720,046
	Val.	159,856	79,902	0	79,954
	Tst.	359	182	0	177
Review	Trn.	34,279	22,259	0	12,056
	Val.	5,925	3,921	0	2,109
	Tst.	7974	5,241	0	2,782
Micro	Trn.	1,360	887	0	473
	Val.	10	6	0	4
	Tst.	340	231	0	108
English	Trn.	152	52	51	48
	Val.	152	49	54	49
	Tst.	2,729	910	906	914
French	Trn.	151	52	51	48
	Val.	152	49	54	49
	Tst.	2,730	910	906	914
Spanish	Trn.	151	52	51	48
	Val.	152	49	54	49
	Tst.	2,730	910	906	914
Italian	Trn.	152	52	51	48
	Val.	152	49	54	49
	Tst.	2,730	910	906	914

Table 1: Dataset statistics. Trn., Val. and Tst. denote training, validation and test sets, respectively. Senti denotes Sentiment140, and Micro denotes Microblog.

Model	# Parameters
Non-pre-trained	1,020,402
Glove	40,402
RoBERTaEmb	87,202
RoBERTa	124,647,170
XLM-RoBERTa & XLM-Twitter	278,045,955
SenticVec	26,102
RoBERTa+SenticVec	124,647,246
XLM-RoBERTa+SenticVec	278,046,069

Table 2: Number of trainable model parameters.

Table 2 shows the parameter sizes of compared models, where SenticVec and Non-pre-trained have the smallest parameter sizes, constituting approximately 3/5 of the parameter size of the Glove model and approximately 3/10 of the parameter size of the RoBERTaEmb model. Conversely, PLMs significantly surpass SenticVec in terms of parameter size by several orders of magnitude.

4.3 Setup

We use Adam optimizer (Kingma and Ba, 2014). We set the learning rate to 1e-3 which decays by 0.5 every 10 epochs. The models are trained with a batch size of 20 for 50 epochs with early stopping based on validation accuracy. We implement $I = 2$ layers of HAN for SenticVec construction. We use spaCy[4] for PoS tagging and lemmatization. We use the 6B GloVe. The hidden dimension for all MLPs is set to 100. Results are averaged from 5 runs.

5 Results

We evaluate our neurosymobic method in conventional supervised learning and cross-domain inference (Section 5.1); cross-lingual inference (Section 5.2) and multilingual learning (Section 5.3). Our ablation analysis (Appendix A) verifies that as the dimensionality of SenticVec increases, our model can achieve better accuracy. We further compare the convergence curves of our method and the baselines (Appendix B), demonstrating the training speed advantage inherent in our method.

5.1 Supervised and Cross-domain Evaluation

For conventional supervised learning evaluation, we use the training and testing sets from the same datasets. As seen in Table 3 (gray), SenticVec achieves comparable performance as RoBERTa but with much fewer learnable parameters (Table 2). When evaluated using F1 score, SenticVec exceeds RoBERTa on the Senti.140 and Review datasets. This compelling evidence substantiates the potential of a shallow neural network to attain performance comparable to a deep neural network-based PLM, through the strategic integration of efficient pragmatic features (the SenticVec knowledge base), alongside syntactic features (PoS tags). Training the SenticVec model uses much less computational resources, given it is merely a 3-layer MLP model. Compared to other embedding-based methods, i.e., non-pre-trained embedding baseline, Glove-based model, and RoBERTaEmb-based model, the advantage of SenticVec is even more prominent. It shows the utility of SenticVec compared to other semantic embeddings, especially in the context that SenticVec representations have a lower dimensionality than the pre-trained embeddings. Finally, RoBERTa+SenticVec achieves the best performance, because the in-domain fine-tuning with deep neural networks together with SenticVec allows the model to learn more complementary domain-specific features. However, this finding does not hold for cross-domain inference, because cross-domain inference may prefer robust general features. Our cross-domain evaluation tasks are conducted using Senti.140, Review, and Microblog. A model is trained with one dataset and evaluated with the rest. Results are shown in Table 3. When using Senti.140 as the training set, SenticVec achieves the best cross-domain performance. Considering that Senti.140 has the biggest training set, we believe that, given sufficient training samples, SenticVec can achieve better results than PLM-based methods. Although RoBERTa+SenticVec is better than SenticVec on Senti.140 testing set, it performs worse than SenticVec during cross-domain evaluation.

[4] https://spacy.io/usage/linguistic-features

Trn	Model	Senti.140 Acc	Senti.140 F1	Review Acc	Review F1	Microblog Acc	Microblog F1	Cr-dm Avg Acc	Cr-dm Avg F1
Senti.140	Non-PT	82.17	83.25	75.48	79.81	65.02	73.66	70.25	76.74
	Glove	85.03	85.71	77.55	81.32	65.79	73.47	71.67	77.40
	RbtEmb	83.01	83.99	76.98	80.95	64.09	72.90	70.54	76.93
	Rbt	86.07	86.56	77.99	81.32	70.90	78.54	74.45	79.93
	SenticVec	84.52	86.86	**83.32**	**87.28**	**73.92**	79.01	**78.62**	**83.15**
	Rbt+SV	**88.39**	**87.90**	79.95	82.94	73.92	**80.49**	76.94	81.72
Review	Non-PT	67.13	72.04	84.81	88.41	60.06	69.93	63.60	70.99
	Glove	78.27	80.20	88.67	91.14	72.79	79.40	75.53	79.80
	RbtEmb	78.27	79.58	88.82	91.44	71.52	78.80	74.90	79.19
	Rbt	85.74	85.36	90.02	92.92	73.92	78.12	79.83	81.74
	SenticVec	84.73	85.16	89.20	94.02	**74.95**	**81.28**	79.84	83.22
	Rbt+SV	**88.50**	**88.89**	**93.10**	**94.71**	74.64	79.84	**81.57**	**84.37**
Microblog	Non-PT	61.56	65.67	62.58	70.85	77.09	83.63	62.07	68.26
	Glove	65.18	71.26	69.55	79.57	83.55	88.15	67.35	75.42
	RbtEmb	61.28	65.85	66.56	76.75	79.57	85.20	63.92	71.30
	Rbt	82.73	84.34	82.51	86.30	85.45	89.55	82.62	85.32
	SenticVec	78.04	**89.01**	**83.66**	**90.26**	84.91	89.12	80.85	**89.63**
	Rbt+SV	**86.85**	87.16	83.70	87.24	**90.19**	**92.83**	**85.26**	87.20

Table 3: Supervised learning and cross-domain inference results. Trn denotes the used training set. The gray denotes conventional supervised learning resutls. Cr-dm Avg denotes the averaged cross-domain inference scores, excluding the gray.

This suggests that the augmented semantic features provided by RoBERTa do not synergize effectively with the syntactic and pragmatic features in sentiment analysis tasks when sufficient training data is available. SenticVec knowledge has a higher utility in general feature representation, compared to the features learned by RoBERTa. This can be explained by the well-known overfitting issue of deep neural networks (Rice et al., 2020) when given a large amount of domain-specific training data. Deep neural networks may overfit in the data distribution of a specific domain. As a result, the learned features struggle to generalize to other domains. Unsurprisingly, SenticVec exceeds all embedding-based methods with large margins.

When Review is used as the training data, SenticVec achieves better average cross-domain performance than RoBERTa by a small margin. RoBERTa+SenticVec yields the best average performance, suggesting that the ensemble strategy is more effective on a middle-sized training set. SenticVec still has significant advantages over other embedding-based methods. Lastly, Microblog is used for training, and Senti.140 and Review for testing. SenticVec obtains the highest F1 scores, and RoBERTa+SenticVec outperforms RoBERTa on both metrics, indicating that the SenticVec knowledge helps with model robustness against domain changes and class imbalance. It also suggests that with about 1,000 training instances, employing deep nets with semantic learning can yield overall superior results compared to our shallow neural network based on syntactic and pragmatic features.

To sum up, compared to the embedding-based methods, SenticVec obtains superior performance across all cross-domain evaluation tasks. It shows that the combination of syntactic and pragmatic information achieves better utilities than using semantic information in shallow neural network-based sentiment analysis. Furthermore, when sufficient training data is available, SenticVec outperforms RoBERTa. With relatively small training data, employing both RoBERTa and SenticVec representations can yield improved accuracy. The robustness of our proposed SenticVec-based method can be attributed to its multi-dimensionality. Rather than combining all pre-training data to compute a single sentiment score for a lexical unit, we uphold the importance of diverse human annotations across various annotation tasks and domains, forming multi-dimensional representations.

We regard each pre-training dataset as a repository of knowledge contributed by a specific group of annotators with their subjective judgments, potentially tailored to specific domains. Given a test set from a separate group of annotators, our framework enables models to judiciously select and learn from the sentiment knowledge aligning best with the domain preferences of the test set. Thus, SenticVec can achieve robust inference for data from other domains. The utility of SenticVec, when trained on a large dataset, holds practical importance for cross-domain inference. Since SenticVec relies exclusively on PoS and pragmatic knowledge as input, it can alleviate the influence of semantic variations stemming from different domains.

Model	Spanish		French		Italian		Average	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
XLM-Rbt	44.78	42.17	45.95	45.72	45.09	44.32	45.27	44.07
XLM-Twt	44.05	40.62	46.06	45.83	44.29	43.48	44.80	43.31
SenticVec	**47.33**	**46.05**	45.88	44.37	**47.07**	**47.07**	**46.76**	**45.83**
XLM+SV	45.03	43.18	**48.75**	**48.89**	45.28	45.07	46.35	45.71

Table 4: Cross-lingual inference results by training with the English dataset. XLM-Rbt denotes XLM-RoBERTa; XLM-Twt denotes XLM-Twitter; XLM+SV denotes XLM-RoBERTa+SenticVec.

The effectiveness of SenticVec is predominantly contingent on the PoS patterns it has learned through training. A large dataset likely encompasses a variety of PoS patterns, thereby enabling effective training of the SenticVec model even though the samples are from the same domain. This could potentially provide insight into an effective approach for fine-tuning neurosymbolic models across different domains: Annotated data should be chosen with the objective of enhancing the diversity of PoS patterns in the fine-tuning dataset. Given the relatively limited vocabulary of PoS tags, a viable strategy involves selecting a subset of data that exhibits a wide range of PoS patterns. Annotators can then focus their efforts on annotating this carefully chosen subset of data. Further investigation into this strategy will be a focal point of our future research.

5.2 Cross-lingual Evaluation

In this section, we conduct cross-lingual inference where models trained with few-shot English data, and tested in Spanish, French, and Italian, using the datasets, developed by Barbieri et al. (2021). Table 4 shows that SenticVec exceeds baselines on average. The slightly inferior performance in French compared to the baselines might be because French is less syntactically similar to English. It indicates that the syntactic and pragmatic patterns learned in English can be applied to other languages to some extent. XLM-RoBERTa+SenticVec outperforms the baselines on all languages, suggesting that sentiment concepts learned from the SenticVec knowledge base transfer well on cross-lingual tasks, hence able to bring extra performance gains to PLMs. To sum up, SenticVec is a strong competitor to fine-tuning-based multilingual PLMs in few-shot multilingual and cross-lingual evaluation tasks, albeit being much more lightweight. This is attributed to the advantage of our novel symbolic representations and shallow neural networks.

Dataset	XLM-Rbt		XLM-Twt		SenticVec		XLM+SV	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
English	53.19	53.43	55.05	52.87	**57.80**	**57.63**	55.33	54.92
Spanish	35.60	26.16	37.22	32.49	**47.44**	**46.85**	41.55	38.85
French	50.37	50.04	37.33	36.71	47.03	47.04	**53.12**	**52.01**
Italian	48.86	48.42	48.75	43.37	50.00	49.66	**50.90**	**50.77**
Average	47.01	44.52	44.59	41.36	**50.57**	**50.30**	50.23	49.14

Table 5: Few-shot multilingual experiment results.

The PoS tags possess a limited vocabulary and the capability to break language barriers, thus applicable across diverse languages. The shallow neural networks can be easily trained with few-shot data and limited computational resources because of the small PoS tag vocabulary and the limited learnable parameters. The model only needs to learn the dependency between unified PoS tags rather than that between words from different languages.

SenticVec knowledge base can provide useful pragmatic knowledge for sentiment analysis in other languages. The sentiment polarities of emotional concepts, e.g., "happiness" and "sadness", are generally consistent across languages. Hence, the sentiment-pragmatic knowledge can be easily broadcasted to other languages via the mappings of multilingual lexicons, delivering robust performance in multilingual sentiment analysis. On the other hand, researchers also identified nuanced semantic differences in emotional words (Zhang et al., 2024). Thus, making appropriate SenticVec adjustments for non-English texts is also recommended.

The robust performance observed in cross-lingual and few-shot sentiment analysis tasks holds practical significance. Leveraging languages with rich labeled data such as English to pre-train a model allows for its subsequent fine-tuning with a limited set of labeled data in the target language. Another notable application involves utilizing a well-trained model for predicting labels in a zero-shot setting for a language that lacks labeled data entirely. We will study them in future work.

5.3 Multilingual Evaluation

Table 5 shows the results of training and inference with the same target languages, i.e., with the English, Spanish, French, and Italian datasets under a few-shot setup, where 152 samples are used for training. As shown, the difference between the baselines' best and worst F1 scores are $\Delta = 27.27\%$ and $\Delta = 20.38\%$ for XLM-RoBERTa and XLM-Twiter, respectively. In contrast, SenticVec achieves an averaged F1 score 5.78% higher than the best baseline XLM-RoBERTa, with $\Delta = 10.78\%$. The higher average F1 score and minimal performance disparity observed across the four languages confirm that our model is more robust in few-shot and multilingual sentiment analysis. Introducing additional XLM-RoBERTa features upon SenticVec knowledge brings extra gains compared to the SenticVec model in some cases, and propels XLM-RoBERTa+SenticVec to exceed the PLM baselines in all target languages.

6 Conclusion

We propose a novel neurosymbolic model and a sentiment-pragmatic knowledge base, SenticVec. SenticVec consists of multi-dimensional sentiment-pragmatic representations, making it human-centric and cross-domain efficient. We propose to use PoS tags together with SenticVec for sentiment analysis. The streamlined PoS vocabulary as input markedly reduces the need for deep neural networks and overcomes language barriers. Consequently, employing a shallow neural network for sentiment analysis in multilingual learning and cross-lingual inference scenarios becomes viable, substantially decreasing computational demands. The proposed framework can be adapted to other text classification tasks with multiple labeled datasets, e.g., emotion detection (Mao et al., 2023b), depression detection (Han et al., 2022), and topic classification (Duong et al., 2023), which will be examined in our future works.

Limitations

The proposed framework have the following limitations. First, the quality of non-English entries in the knowledge base would be affected by the capability of the selected machine translation systems. Second, our proposed model is only tested on Latin alphabetic languages. It might not work as well on languages with more complex segmentation.

Ethics Considerations

The SenticVec knowledge base and involved models are based on public corpora that do not contain private data nor offensive content.

Acknowledgments

This research/project is supported by the Ministry of Education, Singapore under its MOE Academic Research Fund Tier 2 (STEM RIE2025 Award MOE-T2EP20123-0005), as well as supported by Alibaba Group and NTU Singapore.

References

Charu C Aggarwal et al. 2018. Neural networks and deep learning. *Springer*, 10(978):3.

Mohammad Anas, Anam Saiyeda, Shahab Saquib Sohail, Erik Cambria, and Amir Hussain. 2024. Can generative ai models extract deeper sentiments as compared to traditional deep learning algorithms? *IEEE Intelligent Systems*, 39(2):5–10.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of LREC*, page 2200.

Nesrine Bannour, Sahar Ghannay, Aurélie Névéol, and Anne-Laure Ligozat. 2021. Evaluating the carbon footprint of nlp methods: a survey and analysis of existing tools. In *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*, pages 11–21.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2021. Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond. *arXiv preprint arXiv:2104.12250*.

Christin Beck, Hannah Booth, Mennatallah El-Assady, and Miriam Butt. 2020. Representation problems in linguistic annotations: Ambiguity, variation, uncertainty, error and bias. In *14th Linguistic Annotation Workshop (LAW 14)*, pages 60–73.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ B. Altman, Simran Arora, Sydney von Arx, et al. 2021. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1352–1362.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *NeurIPS*, 33:1877–1901.

Erik Cambria, Rui Mao, Melvin Chen, Zhaoxia Wang, and Seng-Beng Ho. 2023. Seven pillars for the future of artificial intelligence. *IEEE Intelligent Systems*, 38(6):62–69.

Erik Cambria, Xulang Zhang, Rui Mao, Melvin Chen, and Kenneth Kwok. 2024. SenticNet 8: Fusing emotion AI and commonsense AI for interpretable, trustworthy, and explainable affective computing. In *Proceedings of International Conference on Human-Computer Interaction (HCII)*.

Yash Chaudhary. 2020. Stock-market sentiment dataset.

Yo Joong Choe, Kyubyong Park, and Dongwoo Kim. 2020. word2word: A collection of bilingual lexicons for 3,564 language pairs. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3036–3045.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Keith Cortis, André Freitas, Tobias Daudert, Manuela Huerlimann, Manel Zarrouk, Siegfried Handschuh, and Brian Davis. 2017. SemEval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. In *Proceedings of SemEval*, page 519.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, page 4171.

Kelvin Du, Frank Xing, Rui Mao, and Erik Cambria. 2023. FinSenticNet: A concept-level lexicon for financial sentiment analysis. In *2023 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 109–114.

Cuc Duong, Vethavikashini Chithrra Raghuram, Amos Lee, Rui Mao, Gianmarco Mengaldo, and Erik Cambria. 2023. Neurosymbolic ai for mining public opinions about wildfires. *Cognitive Computation*.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1(12):2009.

Sooji Han, Rui Mao, and Erik Cambria. 2022. Hierarchical attention network for explainable depression detection on Twitter aided by metaphor concept mappings. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 94–104.

Kai He, Rui Mao, Tieliang Gong, Erik Cambria, and Chen Li. 2022a. JCBIE: A joint continual learning neural network for biomedical information extraction. *BMC Bioinformatics*, 23(549).

Kai He, Rui Mao, Tieliang Gong, Chen Li, and Erik Cambria. 2022b. Meta-based self-training and reweighting for aspect-based sentiment analysis. *IEEE Transactions on Affective Computing*, 14:1731–1742.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9:1735.

Mochamad Nurul Huda. 2023. Bag brands sentiment dataset.

Chihli Hung and Shiuan-Jeng Chen. 2016. Word sense disambiguation based sentiment lexicons for sentiment classification. *Knowledge-Based Systems*, 110:224–232.

Sherif Hussein. 2021. Twitter sentiments dataset. *Mendeley Data, V1*.

Ronald Kemker, Marc McClure, Angelina Abitino, Tyler Hayes, and Christopher Kanan. 2018. Measuring catastrophic forgetting in neural networks. In *Proceedings of AAAI*, volume 32.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Jan Kocoń, Joanna Baran, Marcin Gruza, Arkadiusz Janz, Michał Kajstura, Przemysław Kazienko, Wojciech Korczyński, Piotr Miłkowski, Maciej Piasecki, and Joanna Szołomicka. 2022. Neuro-symbolic models for sentiment analysis. In *International Conference on Computational Science*, pages 667–681.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of ACL*, pages 142–150.

Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796.

Rui Mao, Guanyi Chen, Xulang Zhang, Frank Guerin, and Erik Cambria. 2024a. GPTEval: A survey on assessments of ChatGPT and GPT-4. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7844–7866, Torino, Italia. ELRA and ICCL.

Rui Mao, Kai He, Claudia Beth Ong, Qian Liu, and Erik Cambria. 2024b. MetaPro 2.0: Computational metaphor processing on the effectiveness of anomalous language modeling. In *Findings of the Association for Computational Linguistics: ACL*. Association for Computational Linguistics.

Rui Mao, Kai He, Xulang Zhang, Guanyi Chen, Jinjie Ni, Zonglin Yang, and Erik Cambria. 2023a. A survey on semantic processing techniques. *Information Fusion*, 101:101988.

Rui Mao, Qian Liu, Kai He, Wei Li, and Erik Cambria. 2023b. The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection. *IEEE Transactions on Affective Computing*, 14(3):1743–1753.

Jinjie Ni, Rui Mao, Zonglin Yang, Han Lei, and Erik Cambria. 2023. Finding the pillars of strength for multi-head attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 1, pages 14526—14540.

Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 907–914.

David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2021. Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Vinay Venkatesh Ramasesh, Aitor Lewkowycz, and Ethan Dyer. 2021. Effect of scale on catastrophic forgetting in neural networks. In *International Conference on Learning Representations*.

Leslie Rice, Eric Wong, and Zico Kolter. 2020. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pages 8093–8104. PMLR.

Sofia Serrano and Noah A Smith. 2019. Is attention interpretable? *arXiv preprint arXiv:1906.03731*.

Prerna Singh. 2023. Systematic review of data-centric approaches in artificial intelligence and machine learning. *Data Science and Management*, 6:144.

Peeyush Singhal, Rahee Walambe, Sheela Ramanna, and Ketan Kotecha. 2023. Domain adaptation: Challenges, methods, datasets, and applications. *IEEE Access*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of EMNLP*, pages 1631–1642.

Carlo Strapparava and Alessandro Valitutti. 2004. WordNet-Affect: an affective extension of WordNet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, page 1083.

Chang Wei Tan, Christoph Bergmeir, Francois Petitjean, and Geoffrey I Webb. 2020. News headline sentiment dataset.

Enrica Troiano, Sebastian Padó, and Roman Klinger. 2021. Emotion ratings: how intensity, annotation confidence and agreements are entangled. *arXiv preprint arXiv:2103.01667*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Rui Wang, Xu Tan, Renqian Luo, Tao Qin, and Tie-Yan Liu. 2021a. A survey on low-resource neural machine translation. In *Proceedings of IJCAI*, pages 4636–4643. Survey Track.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021b. Kepler: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39:165–210.

Mengde Yang. 2021. A survey on few-shot learning in natural language processing. In *2021 International Conference on Artificial Intelligence and Electromechanical Automation (AIEA)*, pages 294–297. IEEE.

Weijie Yeo, Ranjan Satapathy, Siow Mong Goh, and Erik Cambria. 2024. How interpretable are reasoning explanations from prompting large language models? In *Proceedings of NAACL*.

Xulang Zhang, Rui Mao, and Erik Cambria. 2023a. A survey on syntactic processing techniques. *Artificial Intelligence Review*, 56(6):5645–5728.

Xulang Zhang, Rui Mao, and Erik Cambria. 2024. Multilingual emotion recognition: Discovering the variations of lexical semantics between languages. In *2024 International Joint Conference on Neural Networks (IJCNN)*.

Xulang Zhang, Rui Mao, Kai He, and Erik Cambria. 2023b. Neuro-symbolic sentiment analysis with dynamic word sense disambiguation. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 8772–8783. Association for Computational Linguistics.

Luyao Zhu, Rui Mao, Erik Cambria, and Bernard J. Jansen. 2024. Neurosymbolic ai for personalized sentiment analysis. In *Proceedings of International Conference on Human-Computer Interaction (HCII)*.

A Ablation Analysis

TrnDim.		Senti.140		Review		Microblog	
		Acc	F1	Acc	F1	Acc	F1
Review	1d	75.49	76.22	83.90	87.51	66.87	73.58
	5d	79.11	80.52	85.74	88.87	71.83	79.37
	10d	78.83	79.57	87.76	90.49	68.11	74.31
	15d	83.84	84.57	87.84	90.45	74.61	78.97
	full	84.73	85.16	89.20	94.02	74.95	81.28

TrnDim.		English		Spanish		French		Italian	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1
English	1d	34.69	22.00	33.63	18.39	33.59	19.85	33.14	17.93
	5d	46.81	46.30	39.56	34.76	39.01	36.24	39.11	37.57
	10d	52.23	51.73	42.71	41.00	37.55	30.13	38.75	34.20
	15d	54.84	55.17	45.75	45.02	38.10	38.20	42.85	42.47
	full	57.80	57.63	47.33	46.05	45.88	44.37	47.07	47.07

Table 6: Ablation analysis by different dimensionalities of SenticVec knowledge base, examined via the SenticVec model.

We conduct ablation analysis to show that the integration of new knowledge, i.e., more labeled datasets, can improve model performance on conventional supervised learning, cross-domain inference, and cross-lingual inference tasks in sentiment analysis. We used the first n dimension of SenticVec and the proposed shallow neural network to examine these tasks, where $n \in \{1, 5, 10, 15, 19\}$. $n = 19$ denotes the full dimension of SenticVec. Table 6 shows that generally as dimensionality increases, the neurosymbolic model achieves higher accuracy and F1 scores across all the examined tasks, among which the full dimensionality setup is the best. Given our method of integrating new knowledge into the knowledge base, which involves expanding the dimensionality of vector representations instead of modifying existing elements, we believe that SenticVec has the potential to alleviate the problem of knowledge forgetting often observed in deep neural networks.

B Convergence Curve

As shown in Figure 4, SenticVec demonstrates superior and quicker convergence on Senti.140 compared to both shallow and deep neural network baselines. Conversely, RoBERTa+SenticVec outperforms RoBERTa in both rate and extent of convergence. This observation highlights the efficacy of the sentiment representations derived from the SenticVec knowledge base, expediting convergence process in both shallow and deep neural networks. Consequently, the lightweight SenticVec neurosymbolic model can be trained with significantly reduced computational resources and subsequent carbon emissions in downstream applications.

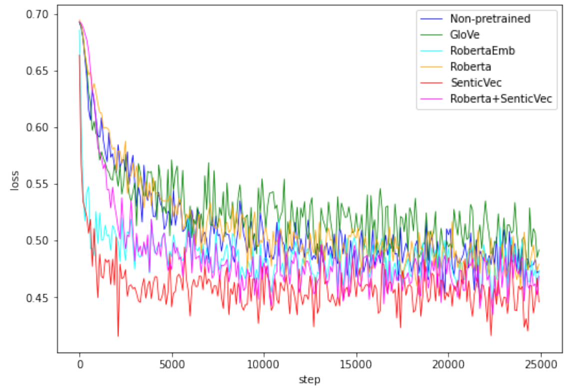

Figure 4: Convergence curves on Senti.140 training set.

C Datasets for Knowledge Base Construction

The datasets used for SenticVec knowledge base construction, their splits, and statistics are shown in Table 7.

Dataset	Origin	Domain	Split	# Samples
sentiment140	Sentiment140 (Go et al., 2009)	Tweet	Train Val	1,440,144 159,856
tweet1	Twitter Tweets Sentiment Dataset[1]	Tweet	Train Val	23,358 4,122
tweet2	Twitter Sentiments Dataset (Hussein, 2021)	Tweet	Train Val	138,523 24,445
tweet3	Sentiment Dataset[2]	Tweet	Train Val	418,762 73,899
airline_tweet	twitter-airline-sentiment[3]	Tweet	Train Val	12,444 2,196
bagbrand	Bag Brands Sentiment Dataset (Huda, 2023)	Product Review	Train Val	2,449 432
sst2	Deeply Moving (Socher et al., 2013)	Movie Review	Train Val	67,349 872
review1	data-reviews-sentiment-analysis[4]	Review	Train Val	6,373 1,125
review2	Generic Sentiment[5]	Review	Train Val	42,464 7,395
yelp_review	Yelp Reviews Sentiment Dataset[6]	Review	Train Val	32,298 5,699
imdb_review	IMDB Large Movie Reviews Sentiment Dataset (Maas et al., 2011)	Review	Train Val	21,247 3,750
stock_tweet1	Stock-Market Sentiment Dataset (Chaudhary, 2020)	Stock	Train Val	4,922 869
headline	Stock-Market Sentiment Dataset[7]	Stock	Train Val	4,119 727
stock_news1	Stock News Sentiment Analysis[8]	Finance	Train Val	92,438 16,313
stock_news2	News Sentiment Analysis for Stock Data[9]	Finance	Train Val	13,228 2,334
bct_tweet	BTC Tweets Sentiment[10]	Finance	Train Val	43,197 7,623
economic_times	Economic Times Sentiment Data[11]	Economics	Train Val	2,132 376
finsentiment	Financial Sentiment Analysis (Malo et al., 2014)	Finance	Train Val	4,966 876
newsheadline	News Popularity in Multiple Social Media Platforms (Tan et al., 2020)	News	Train Val	158,506 27,972

1. kaggle.com/datasets/yasserh/twitter-tweets-sentiment-dataset
2. kaggle.com/datasets/tariqsays/sentiment-dataset-with-1-million-tweets
3. huggingface.co/datasets/osanseviero/twitter-airline-sentiment
4. huggingface.co/datasets/Kaludi/data-reviews-sentiment-analysis
5. kaggle.com/datasets/akgeni/generic-sentiment-multidomain-sentiment-dataset
6. kaggle.com/datasets/thedevastator/yelp-reviews-sentiment-dataset
7. kaggle.com/datasets/ankurzing/sentiment-analysis-for-financial-news
8. kaggle.com/datasets/avisheksood/stock-news-sentiment-analysismassive-dataset
9. kaggle.com/datasets/sidarcidiacono/news-sentiment-analysis-for-stock-data-by-company
10. kaggle.com/datasets/aisolutions353/btc-tweets-sentiment
11. kaggle.com/datasets/rockyjoseph/economic-times-sentiment-data

Table 7: Datasets used for constructing the SenticVec knowledge base.