

Affective awareness in neural sentiment analysis

Rong Xiang^a, Jing Li^{a,*}, Mingyu Wan^c, Jinghang Gu^b, Qin Lu^a, Wenjie Li^a and Chu-Ren Huang^b

^aDepartment of Computing, The Hong Kong Polytechnic University, Hong Kong SAR, China

^bDepartment of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong SAR, China

^cSchool of Foreign Languages, Peking University, Beijing, China

ARTICLE INFO

Keywords:

Deep neural network
Sentiment analysis
Affective knowledge
Sentiment lexicon

ABSTRACT

Sentiment analysis is helpful to bestow ability of understanding human's attitude in texts on artificial intelligence systems. In this area, text sentiment is usually signaled by a few indicative words that convey affective meanings and arouse readers' collective emotions. However, most existing sentiment analysis models have predominantly featured through neural network architectures with end-to-end training manner and limited awareness of affective knowledge, which, as a result, often fails to pinpoint the essential features for sentiment prediction. In this work, we present a novel approach for sentiment analysis by fusing external affective knowledge into neural networks. The affective knowledge is distilled from two sentiment lexicons grounded by two psychological theories, e.g., the Affect Control Theory and word affections in terms of Valence, Arousal, and Dominance. To examine the effects of affective knowledge over sentiment analysis, we conduct cross-dataset and cross-model experiments along with a detailed ablation analysis. Results show that our proposed method outperforms trendy neural networks in all the five benchmarks with consistent and significant improvement (1.4% Accuracy in average). Further discussions demonstrate that all affective attributes exhibit positive effects to model enhancement and our model is robust to the change of lexicon size.

1. Introduction

Sentiment plays a crucial role in shaping people's life decisions, and affects a series of subsequent life choices. Humans tend to act in response to other people's viewpoints in various aspects of life, e.g., where to celebrate Christmas, and whether to get COVID-19 vaccine shot or not, and so forth. To efficiently access people's thoughts and attitudes, a growing number of researchers in Artificial Intelligence (AI) have conducted sentiment analysis based on texts with the hope to help the individuals navigate their decision-making processes. Technically, the task of sentiment analysis is to automatically parse a piece of text and figure out the sentiment polarity of the authors: whether they hold a positive, negative, or neutral attitude towards a particular topic. Sentiment analysis has been regarded as one of the most challenging and essential tasks in artificial intelligence, which strives to help the machines to understand, infer, or even respond to human's emotions [18, 7, 59, 8, 19]. Meanwhile, it will further benefit diverse streams of natural language processing research via providing opinionated features to facilitate downstream tasks [12, 20, 27].


The current research paradigm of sentiment analysis is mainly centering around the use of neural networks to learn features in an automatic manner. This includes the building of various model structure, such as Convolutional Neural Network (CNN) [26], Recursive Auto-Encoders [48], Long-Short Term Memory Network (LSTM) [49], and advancing model training strategies, such as the trendy pre-train and fine-tune paradigm based on transformers [10, 28]. The above models, though achieving promising results, present obvious drawbacks in terms of *interpretability* — they are

unable to explicitly pinpoint the salient words or phrases that link to the sentiment polarity and thus contribute less to reflect human-understandable sentiment components [1, 4, 39]. Moreover, features are learned automatically from large-scale raw data while devaluing a myriad of existing handcrafted external resources, such as linguistic knowledge, cognition-grounded data, and sentiment lexicon, which whereas showed the usefulness in traditional practice [17, 25, 37, 51, 55, 58] and might complement the automatic sentiment features for the AI machines to nurture human senses.

In light of these points, we propose to loan the affective meanings of sentiment words from external resources into neural sentiment analysis frameworks. This allows the model to takes the joint effects of bi-directional feeling between the reader and the writer [44, 53] to make sense of human sentiment. In our intuition, to understand the writers' viewpoints, it is potentially helpful to examine what words they use to voice opinions and how these words resonate with readers' emotions, Figure 1 exemplifies the concepts using two sentences from a movie review collection [41], where S_p refers to the positive sentiment and S_n negative. As can be seen, words with strong affective polarity (“*faith*”, “*sumptuous*” and “*stultifying*”) are strong indicators of the sentiment of the examples. Other affective words, such as “*tale*”, “*love*”, “*betrayal*” and “*revenge*”, as well as some signal words e.g. “*above all*” and “*but*” can be collectively evaluated to determine the overall attitude of the writer.

In this article, we propose to incorporate external **sentiment lexicons** in neural networks for sentiment analysis with human-assigned affective values of various psycho-linguistic aspects (henceforth **affective vectors**). They are manually annotated to reflect how readers' word-level responsive emotions vary along different affective dimensions; for example, “*love*”, usually arousing positive, exciting, and controlling

*Corresponding author

 jing-amelia.li@polyu.edu.hk (J. Li)
ORCID(s): 0000-0002-8044-2284 (J. Li)

S_p : a swashbuckling <i>tale</i> of <i>love</i> , <i>betrayal</i> , <i>revenge</i> and above all , <i>faith</i> .
S_n : visually <i>sumptuous</i> but intellectually <i>stultifying</i> .

Figure 1: Two examples from a movie review collection. S_p exhibits the positive sentiment while S_n negative. Sentiment words are in italic, where red color refers to positive affection while blue negative.

emotions in readers, is likely to be a strong signal of positive sentiment. In previous studies, the helpfulness of leveraging sentiment lexicon has been demonstrated in sentiment analysis task [43, 54, 63]. Nevertheless, most prior efforts focus on leveraging coarse-grained knowledge (i.e., positive and negative words) limited attempts have been made to marry the fine-grained handcrafted affective attributes with the automatic features learned in deep sentiment analysis networks. To fill in the gap, we propose a novel **affective fusion neural network (AFNN)** capable of encoding affective knowledge and highlighting the indicative word patterns with affective attentions. Built upon the success of attention-based neural sentiment analysis [50, 9, 61, 2, 30], we inject affective words as the knowledge seeds into the attention mechanism to capture sentiment indicators, which may enable better interpretability compared with existing attention mechanisms whose features are usually hard to be analyzed by humans [6, 14, 56]. Furthermore, we investigate two alternative training methods to examine how handcrafted affective vectors and automatic semantic representations interact with each other in separate and joint training settings.

To the best of our knowledge, this study is the first to explore the coupled effects of manually-annotated affective vectors and advanced neural networks for sentiment analysis, which somehow involves the implicit engagement of human readers to help sentiment prediction.

To obtain empirical evidence, we carry out a series of experiments based on five benchmark datasets with comprehensive comparisons on different models. The results show that our models outperform the state-of-the-art (SOTA) counterparts by wide margins. For example, RoBERTa with affective attentions achieves 94.2% accuracy on SST-2 movie review dataset which is 1.2% higher than the one without modeling affective knowledge. We also find that a fusion design of both affective vectors and attention mechanism outperforms the alternative that separately train the simple affective augment via feature concatenation and attention mechanism. Next, we analyze the sub-component effects of both the affective vectors EPA and VAD in sentiment prediction and observe that using all factors jointly can achieve the best results. In addition, we test the model’s robustness to the change of sentiment lexicon size and find our model consistently outperforms other methods upon various degrees of knowledge contraction. Finally, case studies on the attention weights show the auto-learning capability of our affection fusion networks, where the attention can highlight the key words indicative of the sentiment polarity.

The rest of this article is organized as follows. Section

2 reviews the previous studies in deep learning based and sentiment lexicon driven sentiment analysis. Section 3 describes the detailed design of our proposed AFNN. Section 4 introduces the experimental setup and the data analysis for the empirical studies, followed by experimental results and discussions in Section 5. Finally, Section 6 concludes this article with some wrapup remarks and look out for some future research directions.

2. Related work

Our work is positioned at the inter-sectional frontiers of sentiment analysis coupling the advances in deep learning and knowledge engineering. In the following, we discuss the previous studies in the respective lines.

2.1. Neural sentiment analysis

The recent decade has witnessed the increasing popularity of neural networks in many language applications including sentiment analysis. We refer to neural sentiment analysis as such a technology focuses on designing and applying deep learning methods to predict the underlying sentiment in texts. In this area of studies, many neural architectures have been proposed and utilized, such as CNN [47], Recursive Neural Networks (ReNN) [48], and Recurrent Neural Networks (RNN) [23]. Among these methods, LSTM (Long short-term memory [49] is one of the most popular designs, which is well-known for its effectiveness in language representation learning. The gated mechanism in the sequence encoder enables LSTM to exhibit the capability of capturing long-term memory and long-distance dependency for sentiment understanding, especially in large-span texts. Based on the aforementioned neural frameworks, attention mechanism is further introduced, which allows models to attend the salient word patterns that may signal sentiment polarity through attention weights [6, 29, 30, 61].

In addition to the advances of model architecture, there exists growing attentions over the development of new training methods — the “pre-train and fine-tune” paradigm has become increasingly popular. Deep language models (e.g., BERT [10], XLNet [60] and RoBERTa [28]) have been proposed to further advance models’ language understanding capability [11, 24]. By pre-training on large-scale unsupervised texts, these models can capture automatic representations, which have advanced forward the cutting-edge results of various tasks, including sentiment analysis [10]. Nevertheless, pre-trained embeddings are in high dimensional space, where it is unclear how the learned feature vectors contribute to the identification of sentiment, not to mention the component effects of different dimensions. We are thus in need of a ‘sensible’ model which is likely to approach the human senses of natural language ‘understanding’ in terms of sentiment ‘analysis’. we propose to employ the affective knowledge captured from human-annotated vectors that reflects dimensional perspectives to words, which can potentially complement the automatic features and provide attributional clues of how to feel like humans. This line of attempts has been under-researched in previous work and will

be comprehensively studied in this article.

2.2. Lexicon-driven sentiment analysis

Another line of research in sentiment analysis has attempted to utilize lexicon-driven approaches by incorporating sentiment lexicon as the prior knowledge through two main strategies. The first way is to craft rules to incorporate sentiment lexicon into the learning results. Wilson et al. [57] and Melville et al. [35] construct a sentiment dictionary for general purposes and demonstrate their helpfulness to linear classifiers. Hutto and Gilbert [22] present a rule-based approach to predict the sentiment scores of the words in lexicon. Andreevskaia and Bergler [5] propose an ensemble system with precision-based weighting for words from the lexicon, which exhibits better accuracy and recall compared with both corpus-based classifiers and lexicon-based systems. Loria [31] presents TextBlob — a well-known python library which uses factorized sentiment terms to infer the sentiment score of the given sentences. Alfrjani et al. [3] exploits DBpedia to calculate the domain specific sentiment polarity, which results in considerable performance gain even to baseline models like Naive Bayes and SVM. These rule-based methods, although easy to use, rely on expertise to customize so that the rules can present obvious constraints in model generalization to new data and domains.

The second way of incorporating sentiment knowledge is to employ vectors or sentiment prior scores to carry lexicon knowledge, which is further injected into the sentiment learning processes. For instance, Teng et al. [54] adopt neural networks to predict the sentiment score of a sentence via exploiting the weighted sum of previously labeled scores of negation words and sentiment words. Qian et al. [43] propose to apply linguistic regularization to sentiment classification based on parse trees, topics, and hierarchical word clusters. Zou et al. [63] adopt a mixed attention mechanism to highlight the roles of sentiment lexicon in the attention layer.

Our model is mostly related with methods leveraging affective knowledge, which usefully boosts sentiment analysis results in previous work. Ma et al. [32] propose Sentic LSTM to incorporate external affective knowledge to a unified hybrid LSTM, exhibiting more than 1% improvement in accuracy. Meškelė and Frasincar [36] present ALDONAr for aspect-based sentiment analysis, where neural attention network is enhanced by incorporating fine-grained knowledge from ontology lexicon. Huang et al. [21] show that sentiment lexicons can be integrated into CNN to allow neural networks to learn both contextual and sentiment representations. Among these methods, attention weights are most widely adopted, attributed to its capability to capture indicative patterns in the local contexts. However, the computation to weigh each word by previous attention requires complicated manipulation over matrices in network layers, which is hence challenging to train from scratch, especially for long sentences. Being different from previous approaches, we employ affective lexicons and their quantitative affective values to guide the attention learning and improve the training

efficiency. Moreover, the affective knowledge in existing models is often coarse-grained, i.e. indicating only binary sentiment polarity, which might not comprehensively reflect people’s understanding of affective meanings.

To mitigate those problems, Xiang et al. [58] examine the effects of fine-grained knowledge by taking annotated sentiment vectors from ACT as the attention weights and demonstrate their helpfulness in neural sentiment analysis. Nevertheless, that work focused on the affective words with annotations and may lead to over-fitting while training. On the contrary, our proposed model explores the interactions between words with affective annotation and others captured by automatic feature learning, which can lower the sensitivity to affective lexicon and exhibit better generalization ability. The details will be further discussed in Section 5.

3. Methodology

In this section, we describe how our AFNN takes the advantage of affective awareness from the handcrafted sentiment lexicons in the neural sentiment classification framework. Specifically, the affective values are integrated into the neural networks using affective attention mechanism. Furthermore, the canonical and affective representation learning branches are fused in a unified loss function to allow joint training. Our intuition is that employing external knowledge may explicitly highlight those words with cognition grounded dimensional information, which may offer the potential benefits to the learning of sentiment representations in neural networks. Detail of AFNN is elaborated in following subsections.

3.1. Preliminaries

We first describe the external resources we employ to capture affective knowledge, followed by the description for input and output.

Affective knowledge. In this article, we consider two external affective knowledge resources, one covers the word annotations in EPA (Evaluation, Potency, and Activity) from **Affective Control Theory (ACT)** [46] and the other is labeled in **Valence, Arousal, and Dominance (VAD)** [40, 45]. In both of them, annotators manually curate affective vectors reflecting readers’ collective emotions to a word over three distinguishing dimensions.

In ACT, people are assumed to maintain common perceptions, where the shared “fundamental” sentiments over a word are separately measured in **Evaluation** (sentiment polarity), **Potency** (affective power), and **Activity** (active degree) (henceforth **EPA**). For instance, the EPA vector of the word “mother” is [2.74, 2.04, 0.67], corresponding to {quite positive}, {quite powerful}, and {slightly active}, respectively. EPA collection adopted in our method is provided by Heise [15] which covers the annotations for the most commonly-used 5,000 English sentiment words.¹

¹http://www.indiana.edu/~socpsy/public_files/EnglishWords_EPAs.xlsx

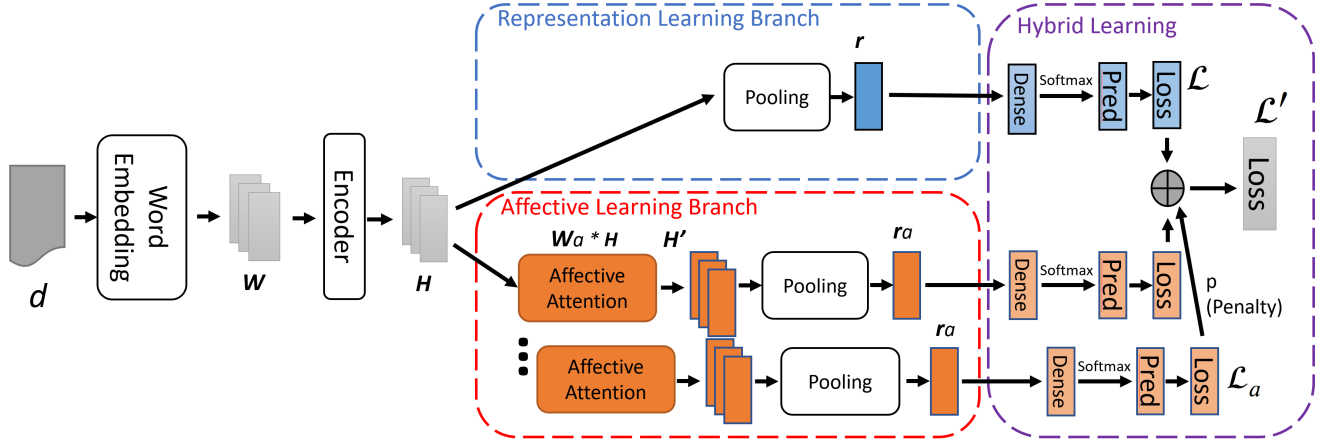


Figure 2: The architecture of AFNN. Blue box shows a typical learning process of deep learning while red box describes how affective learning part grants the capability of affective awareness. They separately handle the task, whose training losses are added with penalty weights for hybrid learning, exhibited in purple box).

VAD lexicon exhibits similar essence, where the affective values are labeled in valence (sentiment polarity), arousal (active degree), and dominance (affective power). For example, the VAD vector of “mother” is $[0.931, 0.408, 0.725]$. These VAD values are provided as a collection with more than 20K English words [38].²

Despite of the slight difference in the values of EPA and VAD in terms of scale (the former ranges from -5.00 to 5.00 while the latter from 0.00 to 1.00), both knowledge resources are measured in three separate dimensions of numerical values in the continuous space. Therefore, we introduce a mapping function to unify the range of EPA and VAD and explore their marriage in affective representation. Concretely, we concatenate the word-level affective vectors from EPA and VAD into a 6-dimensional affective vector, where the value of each dimension is denoted as $v_j, j \in [1, 6]$ corresponding to Evaluation, Potency, Activity, Valence, Arousal, and Dominance, respectively.

Input and output. The goal of sentiment analysis is to assign an affective label for a piece of input texts, reflecting the writers’ attitude. The label types can either be binary for polarity indication or numerical for both polarity and strength. Let \mathbf{D} denote a collection of documents for sentiment classification. Each document $d \in \mathbf{D}$ is first tokenized into a word sequence with maximum length n , then the word embeddings \mathbf{w}^i of the sequence are jointly employed to represent the document $d = \{\mathbf{w}^1, \mathbf{w}^2, \dots, \mathbf{w}^i, \dots, \mathbf{w}^n\} (i \in 1, 2, \dots, n)$.

3.2. Affective fusion neural network

The previous work [58] has attested the usefulness to explicitly encode EPA attentive vectors in attention weights. Nevertheless, it only focuses on affective terms with annotations and ignore the potential sentiment reflection ability of other words, which may weaken the generalization ability of the model and result in the overfitting problem. Taken ad-

vantages of deep neural networks in learning latent semantics in texts, our proposed model AFNN is derived to fuse explicit affective vectors from human annotations and implicit text embeddings from automatic feature learning. As indicated in its architecture in Figure 2, the joint effects are explored with a novel affective attention mechanism, which can be generally applied to many alternatives of neural network models, such as CNN, LSTM, and transformers.

Affective attention weights. To leverage affective annotations, we first mark the affective terms in a dataset before feeding them into deep neural networks. For each recognized affective term, the affective values are indexed from EPA and VAD resources through a linear transformation, which is defined as follows:

$$w_{aj} = 1 + \alpha_j * |v_j| \quad (1)$$

Where α_j stands for a non-negative parameter as the amplification to reflect the influence of the j -th affective dimension for sentiment prediction. Then, w_{aj} is taken as the attention weights, in aware of the affective knowledge, which will later contribute to the learning of the word-level representations. For those words absent in the affective lexicon, we set $w_{aj} = 1$, ignoring the effects from affective knowledge and remaining their semantics from word embeddings. Afterwards, we represent the concatenated affective weights in a vector $\mathbf{w}_a = \{w_{a1}, w_{a2}, \dots, w_{a6}\}$, which carries word-level affection awareness over various dimensions.

Incorporating automatic features. To inject the affective knowledge into the neural framework, we apply a word embedding layer and a pre-trained language model (e.g., BERT) to encode the input document d . The encoded word vectors are then separately fed into two main branches — regular representation learning and affective learning, which are shown in the blue box and red box in Figure 2, respectively.

²<https://saifmohammad.com/WebPages/nrc-vad.html>

In the regular representation learning branch, the document representation vector \mathbf{r} is directly generated through the pooling operation over the involved word embeddings in d . For affective learning, each word embedding $\mathbf{w}^i \in d$ is first factorized into 6 embedding factors, each is initialized by \mathbf{w}^i yet will be adapted to capture affective semantics over each affective dimension (E,P,A,V,A,D) through training. Their weighted sum with affective attention weights (defined in Eq. 1) will hence be used to represent the affective-aware word semantics. Concretely, supposing \mathbf{H} is the matrix holding the concatenated factorization results of \mathbf{w}^i , then the representation captured by the affective attention is computed in the following formula:

$$\mathbf{H}' = \mathbf{w}_a * \mathbf{H} \quad (2)$$

Recall that \mathbf{w}_a is the concatenation of the affective attention weights. As can be seen, besides the potential benefits in effectiveness, our attention design can also exhibit theoretically better efficiency in training and inference. This is because \mathbf{w}_a is captured in lexicon-level rather than document-level and generated through a linear transformation unrelated with the document size, so it can run in constant time. On the contrary, regular attention handles the weights via matrix operations on document level, which requires $O(n)$ costing time given n length documents.

After obtaining the attended representations on word level (\mathbf{H}'), we adopt the average pooling for them to produce document representation \mathbf{r}_a , which integrates both the affective knowledge and the automatic learned embeddings for sentiment prediction, which is presented in the following.

Hybrid learning. In the final classification layer, we should fuse the features from both the regular representations learning branch and affective learning branch. To that end, the two branches first handle the sentiment classification tasks separately; in other words, the learned document representations (i.e., the average pooling results) from each branch are first fed into the dense layers with softmax activation to yield the sentiment prediction. The branch training are both based on the cross-entropy loss, where \mathcal{L} is the loss to train representation learning branch while \mathcal{L}_a is that for affective learning branch.

To further allow the collaborations of two branches, we produce the final loss \mathcal{L}' by trading off the individual effects of regular representation learning (\mathcal{L}) and affective learning (\mathcal{L}_a) with the following formula:

$$\mathcal{L}' = \mathcal{L} + \sum (p_j * \mathcal{L}_{aj}) \quad (3)$$

where \mathcal{L}_{aj} indicates loss derived from the j -th dimension of affective attributes and p_j is the penalty weight to balance the contributions from each constituent of the cumulative loss.

4. Datasets and baseline systems

Experiments are conducted on five datasets. The performance of the proposed method is compared against a series

of commonly used baseline methods as well as SOTA transformer based methods.

4.1. Benchmark datasets

For experiments, the models are investigated on five publicly available and widely adopted benchmark datasets in English: three are from movie review domains while two from social media. The movie review datasets are **SST-2** [48], **MR** [41], and **IMDB** [33], where each review is associated with a binary sentiment label indicating the polarity of *positive* or *negative*. In particular, for **IMDB** dataset, the reviews are collected from IMDb website, where the texts are relatively long and the polarity labels are fully balanced with the ratio of 1:1.³ For the social media datasets, they are both collected from Twitter, where one (named as **Twitter**⁴) concerns diverse range of topics, such as news, public events and daily life, and the other contains customer conversation messages from Twitter in February 2015 from six major American airlines, which is therefore named as **AirRecord**⁵. AirRecord dataset is annotated with three labels (*positive*, *negative*, and *neutral*), which is different from the other four datasets with binary labels.

Table 1

Statistics of the five benchmark datasets. N_{train} : number of training instances. N_{test} : number of testing instances. L : average instance length. N_{voc} : size of vocabulary. C : number of labels.

Dataset	N_{train}	N_{test}	L	N_{voc}	C
SST-2	6,920	1,821	19	16,185	2
MR	9,595	1,067	20	18,765	2
IMDB	22,500	2,500	260	184,885	2
Twitter	89,989	9,999	14	183,645	2
AirRecord	13,172	1,464	18	30,166	3

Table 1 shows the statistics of the five datasets for experiments. They exhibit different characteristics, allowing the model evaluations in various scenarios. For example, Twitter dataset has the largest scale while each instance therein shows the shortest length on average, which may possibly result in data sparsity using feature learning. In comparison, texts in IMDB are very long, showing another challenge to encode rich contexts with complicated structure.

For model evaluations, we follow the practice of original papers to segment the train/test data. For the other datasets without clear instructions from the data suppliers, 10% of the total instances are randomly sampled to for test.

4.2. Affective lexicon analysis

Before experimental investigation, we conducted a preliminary analysis of the affective lexicons (ACT and VAD) employed for affective learning. Figure 3 shows the histograms of three affective measures for both EPA and VAD

³<https://www.kaggle.com/iarunava/imdb-movie-reviews-dataset>

⁴<https://www.kaggle.com/c/twitter-sentiment-analysis2>

⁵<https://www.kaggle.com/crowdflower/twitter-airline-sentiment/home/>

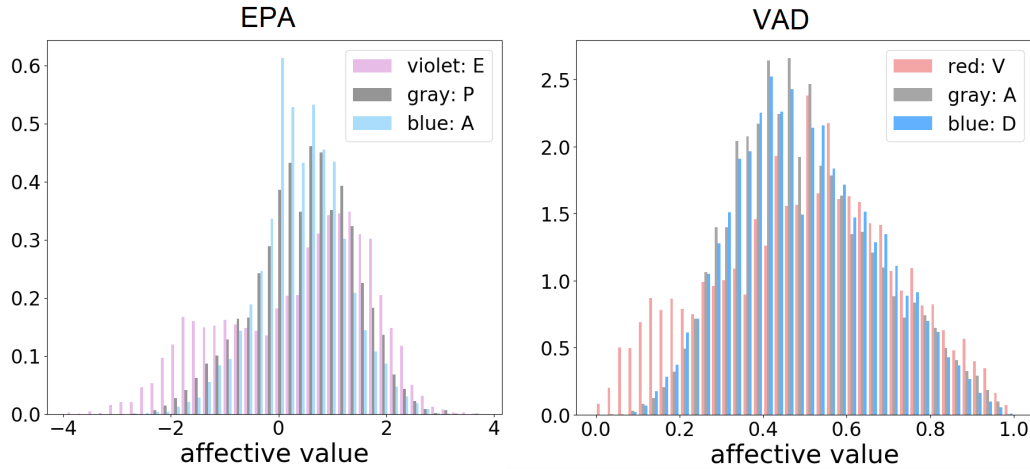


Figure 3: Histograms showing the distribution EPA and VAD values over words. X-axis shows the value in an affective dimension while y-axis displays the probability density of words.

over words from the two affective lexicons. As shown in the figure, none of the three measures exhibits a balanced distribution in EPA and VAD. EPA components are overall right-skewed while Arousal and Domination are slightly left-skewed. Evaluation of EPA and Valence of VAD are the most evenly distributed amongst all, and their distributions are quite similar. Notably, they both have two peaks scattered at the positive axis and the negative axis, which are apparently different from other components. One possible reason is that they both reflect polarity-related attributes in word affection. Potency, Activity, Arousal, and Domination generally follow Gaussian distribution. Since the majority of these four affective values fall near the central area, the affective evidence they provide might be less significant compared with Evaluation and Valence.

Table 2

Proportion of sentiment lexicon in benchmark datasets. EPA and VAD indicate proportion of affection lexicon over all words in the benchmark dataset.

Dataset	EPA(%)	VAD(%)
SST-2	20.0	30.8
MR	20.2	30.5
IMDB	21.2	24.1
Twitter	20.2	20.8
AirRecord	18.1	22.2

To further examine the possible effects of the affective lexicons over the experimental datasets, Table 2 shows the proportion of words from benchmark datasets that have affective annotations in either EPA or VAD lexicon. In general, VAD values are available for higher proportion of words, whereas over 2/3 words lack annotations in prior knowledge and their affection has to be determined via training. We also observe that movie reviews (SST-2, MR, and IMDB) exhibit larger proportion of overlapping words while the percentage in Twitter and AirRecord is relatively lower. This possi-

bly because movie review writers might tend to express their sentiment in an explicit way (using words to clearly indicate their altitude) while the opinions on social media might be more implicit owing to the informal language styles therein.

Then we analyze more on the affection lexicon via demonstrating the proportion of annotated tokens in terms of their part-of-speech (POS) tags in Figure 4. StanfordCoreNLP [34] is adopted for POS tagging. As can be seen, movie review and social media texts exhibit different statistics. The former presents consistent distributions in SST-2, MR, and IMDB, where noun lexicons are most frequently used in all five benchmarks while adjective/adverb terms are least utilized. For the latter, the patterns are hard to be summarized, which again indicates the noisiness of social media texts and the challenges to capture affective senses from them.

4.3. Baseline systems and comparison settings

To examine the effects of affective knowledge in experimental comparisons, three groups of methods are evaluated. The first group includes baseline methods without affective knowledge in both training and test. The baseline systems in this group and their main settings are described as follows:

- **SVM** takes the average of word embeddings of Glove [42] to represent the document vectors.
- **CNN** uses a convolution layer to capture features for adjacent words [26].
- **LSTM** is a popular RNN architecture with a gated mechanism [16] to encode longer contexts.
- **BiLSTM** concatenates the bidirectional encoding results from LSTM to model the word semantics from both left and right neighboring words [13].
- **BiLSTM+AT** is BiLSTM with attention mechanism, designed to combine strengths of both BiLSTM and attention mechanism [62].
- **BERT** is a bidirectional transformer encoder that models contexts through pre-trained language models. [10].
- **RoBERTa** is a refined version of BERT with new self-training tasks for better context modeling [28].

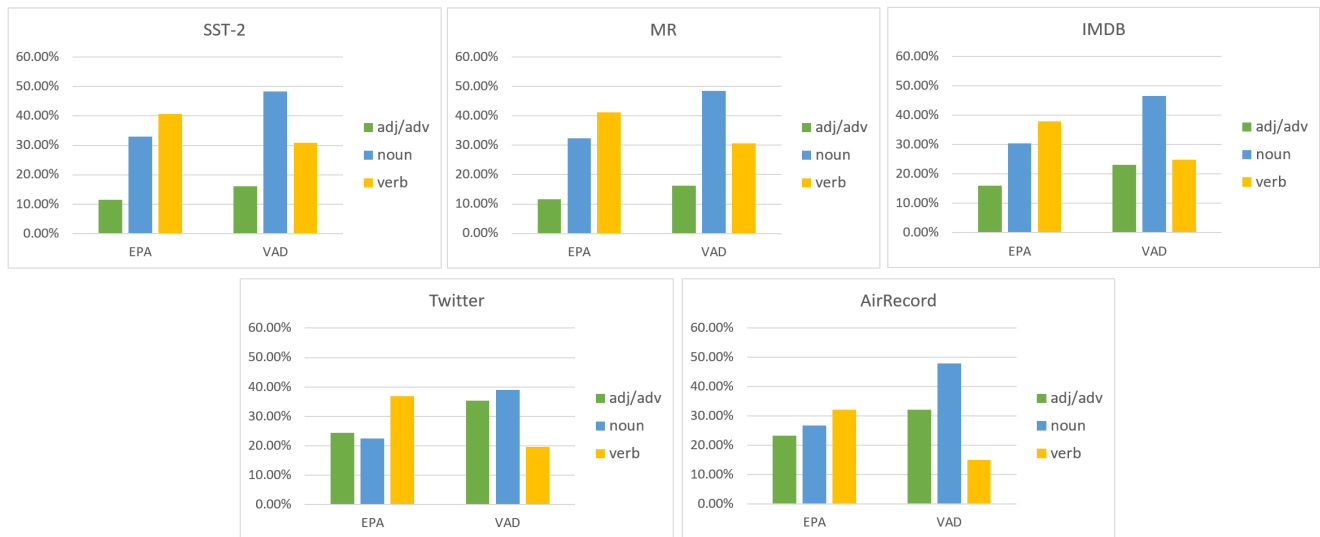


Figure 4: POS proportions of sentiment lexicons over all words in the benchmark datasets.

For models without pre-trained language models, we set the size of word embeddings to 300 and initialized them with pre-trained Glove vectors [42], while the dimension for BERT and RoBERTa embedding is 768. The learning rate is set to $1e-5$ to fine-tune BERT and RoBERTa compared with $5e-4$ for others with pre-trained word embeddings. All the models are trained for 3 epochs with dropout set to 0.1 and batch size 32.⁶ In the evaluation, we report the average accuracy over 5 random seeds for model initialization.

The second group includes a number of methods using affection lexicon which is coarse-grained knowledge, e.g., positive and negative sentiment words. The baseline systems in affection driven group are re-implemented and tested on all benchmark datasets.

- **SSWE** is an SVM-based model using sentiment specific word embedding [52].
- **VADER** (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool [22].
- **TextBlob** is a Python library for processing textual data which provides API for sentiment analysis [31].
- **BiLSTM+SL** incorporates sentiment scores of lexicon into BiLSTM with loss function bias [54].
- **BiLSTM+LBSA** highlights sentiment lexicon with a gold attention mapping in training processes [63].

To further examine the modules of AFNN, we consider its two variant approaches in comparison and form the third group, which incorporates affective knowledge from EPA and VAD. Since both EPA and VAD values are vectors with multiple dimensions, it is straightforward to directly employ them as the sentiment prediction features. This one simple yet effective solution is to use the affective values as augmented lexical features and we hence design the compari-

⁶The batch size for IMDB is set to 6 (instead of 32 done in other datasets) due to the limitation of GPU memory for handling very long documents (as shown in Table 1)

son framework of **Affective Augmented Neural Network (AugNN)**, as shown in Figure 5.

In addition, we include an previous model to the group **Affective Attention Neural Network (AANN)** [58], which adopts EPA and VAD values as the affective attention weight, without the hybrid learning processes with automatic embeddings. Its workflow is shown in Figure 6, where EPA or VAD are projected to a unified value before integrating it into a deep learning model.

To compare various strategies to incorporate EPA and VAD affective vectors, we test AugNN (by simple concatenation), AANN (as attention weights), and AFNN (via affective learning and hybrid learning) based on a number of neural networks, including SOTA models with pre-trained BERT and RoBERTa, and BiLSTM baseline with word embeddings. In implementation, we reproduced the results and adopted the best reported parameters for all benchmarks to allow comparable results. For AFNN, the penalty weights p_j are set to [0.2, 0.3, 0.3, 0.2, 0.3, 0.3] for EPA and VAD dimensions, respectively.

5. Experimental results and discussions

In this section, we first systematically evaluate the performance of our approach on the benchmark corpora, followed by the substantial analyses over the effects of affective knowledge. Specifically, we present the comparison results with the baselines and SOTA models in Section 5.1. Then, we compare AFNN with other alternatives to encode affective knowledge in Section 5.2. Next, the effects of varying affective dimensions in ACT and VAD are analyzed in Section 5.3. Afterwards, the sensitivity of AFNN towards affective lexicon is discussed in Section 5.4. Lastly, two example cases are discussed in Section 5.5 to qualitatively analyze the outputs and interpret the advantages of AFNN.

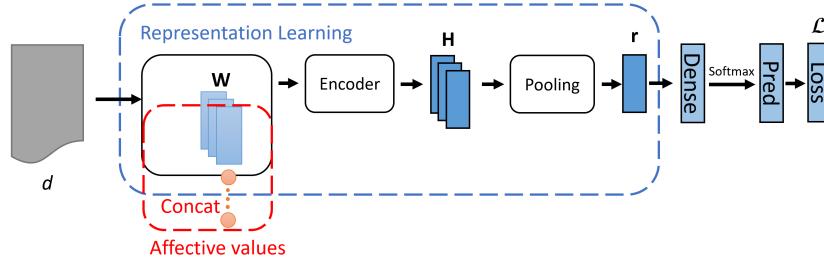


Figure 5: Framework of the comparison model AugNN. Affective vectors are directly adopted as the features and concatenated with automatic features for sentiment prediction.

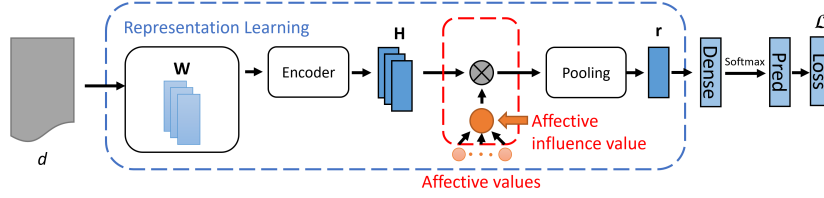


Figure 6: Framework of the comparison model AANN. It directly takes the affective values as the attention weights without the hybrid learning process.

5.1. Main comparison results

In this section, the overall performance is evaluated on the five benchmark datasets. Table 3 reports the classification accuracy grouped in none-affective (NAFF) methods, affective driven models (ADRV), and our variants with affective awareness (AAWR), including propose framework AFNN. The methods using AugNN, AANN and AFNN are suffixed as +Aug, +AA and +AF for better display.

In the NAFF group (on the top), SVM performs the worst, indicating the limitations of using shallow features of average embeddings to capture context information to predict the document-level sentiment. In contrast, neural models exhibit much better accuracy, among which, CNN performs the worst for long texts (e.g., IMDB). A possible explanation is that introducing a convolution window may focus on exploring local contexts with neighboring word occurrence patterns, whereas the fixed window size may mistakenly segment the consistent semantics and thus result in the incorrect understanding of the global context. LSTM-based models, on the other hand, can manage to track long-term dependency and partially solve the vanishing gradient problem. Therefore, they present better results than CNN and SVM in IMDB. However, interestingly, BiLSTM and BiLSTM+AT cannot provide much performance gain over LSTM, especially in SST-2 and MR, which is probably owing to the relatively smaller training sets (as shown in Table 1) to learn a large set of parameters. In addition, we observe the much better accuracy achieved by BERT and RoBERTa, possibly benefited from their language understanding ability gained from the pre-training on large-scale corpora.

In the group ADRV (in the middle), SSWE, which utilizes SVM and affective lexicon, results with 1% accuracy improvement over SVM, which shows the usefulness to incorporate external sentiment lexicon. For VADER and TextBlob (both taking sentiment lexicon as the dictionary), they

Table 3

Sentiment classification accuracy. Models from top to bottom are grouped in NAFF (w/o affective modeling), ADRV (existing affective-driven models), and AAWR (our variants with affective awareness). Boldface scores indicate the best results for each group.

	SST-2	MR	IMDB	Twitter	AirRecord
SVM	71.8	70.2	69.5	61.4	70.5
CNN	81.4	79.1	76.1	70.3	76.8
LSTM	80.2	77.0	80.3	74.7	80.5
BiLSTM	80.7	77.6	79.7	75.6	80.7
BiLSTM+AT	79.5	77.9	80.5	75.9	81.3
BERT	91.3	87.1	88.1	82.0	83.2
RoBERTa	93.0	90.3	89.1	83.3	84.3
SSWE	73.2	71.1	70.3	62.8	71.1
VADER	83.6	82.5	75.4	73.5	78.3
TextBlob	84.0	82.7	75.9	73.6	78.3
BiLSTM+SL	82.1	78.9	81.8	76.1	81.4
BiLSTM+LBSA	81.8	79.1	81.3	76.7	81.8
BiLSTM+Aug	80.9	77.8	80.0	75.9	80.8
BiLSTM+AA	81.1	78.6	82.2	76.6	82.2
BiLSTM+AF	81.7	79.2	82.4	76.9	82.4
BERT+AF	92.2	88.4	89.5	83.2	84.2
RoBERTa+AF	94.2	91.4	90.8	84.5	85.4

achieve fine results in SST-2 and MR while we observe the opposite for the rest three datasets, whose data is more complicated or noisier and hence are more likely to render the appearance of multiple sentiment words indicating conflicting polarities. BiLSTM+SL and BiLSTM+LBSA achieve the best performance in this group, probably owing to the involvement of sentiment lexicons in training, either as additional sentiment prior or pre-defined attention weights. As the result, the lexicon knowledge is jointly explored with other neural modules, which may help transfer the prior knowledge to learn semantics of the words absent in the sentiment lexicon.

For Group AAWR (on the bottom), we compare its three modules (Aug, AA, and AF) over BiLSTM baseline and further incorporate AF module into BERT and RoBERTa to examine its collaboration results with varying deep semantic representations. It can be observed that BiLSTM+Aug exhibits very limited improvement over BiLSTM, implying that feature concatenation cannot guarantee good affective modeling. One possible reason is the unbalanced feature dimensions: Glove word embeddings are 300-dimensional vectors, which is much larger than the affective vectors, which is up to 6 considering both EPA and VAD. For BiLSTM+AA, its results are close to BiLSTM+SL and BiLSTM+LBSA, which shows that attention can be a good alternative to exploit affective knowledge, though better design may be needed to further surpass the previous SOTA. Moreover, better accuracy is presented by BiLSTM+AF than BiLSTM+AA; it suggests the effectiveness of affective fusion mechanism to model the interactions between affective knowledge (from EPA and VAD) and automatic features (captured by deep learning). Likewise, the design of affective fusion boosts BERT and RoBERTa based classifiers, which suggests that the handcrafted affective knowledge can provide further benefits to the representations learned by the BERT family.

5.2. Further discussions on AFNN variants

The preliminary results in Table 3 has shown the potential superiority of AFNN in exploiting affective knowledge. Furthermore, the comprehensive results of AugNN, AANN, and AFNN are discussed based on four popular neural networks: two LSTM variants (LSTM [16] and bidirectional LSTM (BiLSTM) [62]) and two classifiers based on pre-trained language models (BERT [10] and RoBERTa [28]). All these models are first implemented individually and then extended with the AugNN, AANN and AFNN modules correspondingly. They are associated with the most effective affective values, where the selection of values will be further discussed in Section 5.3. Table 4 reports the accuracy of the vanilla models in comparison to their counterparts gaining affective awareness with the three alternatives in AAWR.

In general, affective awareness methods show better results than their ablations without affective knowledge modeling, and AFNN consistently outperforms other comparison models when combined with different bases. AugNN exhibits the smallest performance gain and sometimes even results in worse results, which suggests that simple feature concatenation (augmentation) may not allow the model to effectively couple affective knowledge and deep learning features, because of the different semantic spaces they are in. For AANN, it can leverage affective vectors as the prior attention amplification and results in better accuracy than the straightforward AugNN schema, which may suggests an alternative to merge affective vectors and latent word embeddings. Nonetheless, AFNNs obtain further performance gain by effectively capture how affective knowledge and representation learning interact with each other to work together for sentiment indication.

In comparison of different datasets, AFNN obtains the

Table 4

Accuracy comparison of base neural models and its counterparts with affective awareness from Aug, AA, and our proposed method AF modules. AugNN, AANN, AFNN exhibit 0.17%, 1.07% and 1.39%, respectively, over all methods across the five benchmarks. AFNN significantly outperforms base models with the absolute performance gain of 1.54%, 1.66%, 1.16%, and 1.26% over LSTM, BiLSTM, BERT, and RoBERTa (p -value < 0.001, paired t-test).

	SST-2	MR	IMDB	Twitter	AirRecord
LSTM	80.2	77.0	80.3	74.7	80.5
+Aug	80.5	77.2	80.5	74.9	80.8
+AA	81.1	77.8	81.8	75.9	82.3
+AF	81.3	78.1	82.2	76.3	82.5
BiLSTM	80.7	77.6	79.7	75.6	80.7
+Aug	80.9	77.8	80.0	75.9	80.8
+AA	81.1	78.6	82.2	76.6	82.2
+AF	81.7	79.2	82.4	76.9	82.4
BERT	91.3	87.1	88.1	82.0	83.2
+Aug	91.3	87.2	88.0	82.2	83.4
+AA	92.0	88.0	89.1	83.1	83.8
+AF	92.2	88.4	89.5	83.2	84.2
RoBERTa	93.0	90.3	89.1	83.3	84.3
+Aug	93.1	90.2	89.2	83.4	84.4
+AA	93.8	91.1	90.3	84.1	85.1
+AF	94.2	91.4	90.8	84.5	85.4

average performance gains of 0.63%, 0.75%, 1.20%, 0.85%, and 0.93%, respectively, over SST-2, MR, IMDB, Twitter, and AirRecord datasets, where the latter three exhibit slightly larger boost. This could be attributed to the challenges to automatically learn essential features from very long contexts with complicated structure (IMDB) or very short ones with data sparsity (Twitter and AirRecord). In these cases, human-annotated affective knowledge could be helpful to guide the neural representation learning modules to produce meaningful embeddings for sentiment analysis.

5.3. Feature analyses for affective knowledge

Previous discussions concern how affective knowledge obtained from EPA and VAD helps to boost the overall performance. In this section, we probe into the subordinate effects of the all 6 affective components and study the contribution of each affective dimension.

The following feature analyses further investigates the effect of affective fusion for enhancing four neural networks through 6 compositional dimensions. In ACT, the three components {E, P, A} are used as affective information to characterize an affective-related event while {V, A, D} is designed to model affective factors for VAD lexicon. The performance gain with respect to LSTM, BiLSTM, BERT and RoBERTa are averaged across five datasets. Figure 7 depicts the average improvement for individual and combined affective feature(s) for EPA and VAD, separately.

For single dimension application, Evaluation and Valence are more effective than other measures in EPA and VAD, respectively. Evaluation seems to be the most dominant dimension, exhibiting consistently leading performance in com-

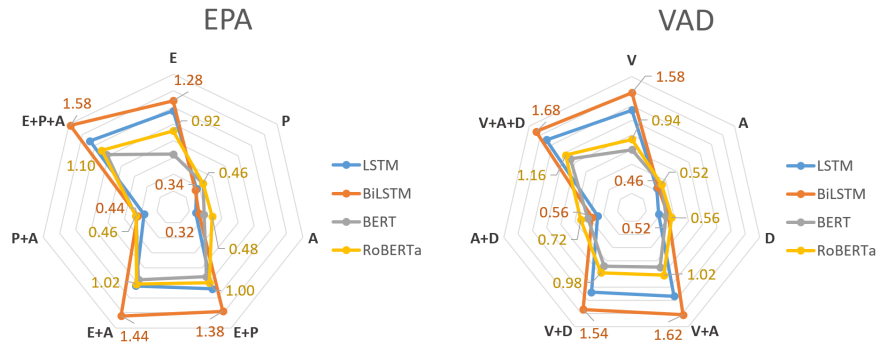


Figure 7: Ablation studies of EPA (left) and VAD (right). The numbers present the performance gain of BiLSTM and RoBERTa, in orange and yellow, respectively.

combination with either Potency, Activity or both; similar observations are drawn for Valence from the VAD radar graph. This result is intuitive since these two measures indicate word-level polarity by design, whereas the others are supplemental and descriptive information. Although the improvements of using Potency, Activity, Arousal and Dominance are quite marginal, they can still potentially contribute as additional knowledge for distributing attention weight. In addition, it can be found that BERT and RoBERTa result with more accuracy gain using Arousal and Dominance than Potency and Activity, probably because VAD lexicon presents more word annotations (as discussed in Section 3.1), hence enabling better marriage with the large-scale pre-trained language model.

Performance gain from one single dimension is smaller compared with its counterparts with any forms of dimension combination, involving either 2 or 3 attributes. Furthermore, using all three affective components result in the best performance. This implies that introducing richer external knowledge could provide more affective information for sentiment identification.

In terms of different prototypes, LSTM-based classifiers gain more benefit from affective knowledge than BERT and RoBERTa. One possible reason is pre-trained language models can gather more information from external language resources. Nevertheless, in most scenarios, BiLSTM and RoBERTa outperform LSTM and BERT respectively but the gap in between are both marginal.

5.4. Sensitivity to affective lexicon

Affective-enhanced methods usually heavily rely on sentiment lexicon, causing problems of knowledge-biases and domain-constraints. To examine AFNN's capability against such problems, four affective driven solutions are investigated under three conditions: (i) *scarce* (instances containing no more than one sentiment lexicon); (ii) 10% instances with least sentiment lexicons; (iii) 10% instances with most sentiment lexicons. We first sample the test instances fitting one of the above three conditions and examine the accuracy of VADER, BiLSTM+LBSA, BiLSTM, and RoBERTa+AF

over them. The results are demonstrated in Figure 8.⁷

The overall performance rank is RoBERTa+AF, BiLSTM+AF, BiLSTM+LBSA, and VADER, from best to worst. All classifiers exhibit increasing trend given higher sentiment lexicon ratio. This means that test samples with more words having affective annotations will better gain the affective awareness from prior knowledge. Across multiple classifiers, VADER is most sensitive to sentiment lexicon ratio, indicating its heavy reliance on the availability of annotated sentiment terms in test instances. In particular, for the scarce instances containing merely no sentiment lexicon, VADER performs rather poorly with the accuracy of around 50% to 60%, barely outperforming random guesses.

If we compare BiLSTM+LBSA and BiLSTM+AF, it is seen that the latter exhibits less error cases in the scarce and least 10% group, and RoBERTa+AF further higher the performance gain in these samples with sparse annotations. This demonstrates the ability of our AFNN to handle inputs with sparse sentiment lexicon, possibly because its hybrid learning processes (leveraging both regular representation learning and affective learning) may generalize the hand-crafted knowledge to other words without annotations. On the contrary, BiLSTM+LBSA uses sentiment lexicon alone as the gold attention, which is therefore outperformed by BiLSTM+AF. Another advantage of AFNN comes from the capability to encode fine-grained affective annotations from EPA and VAD, which provide richer senses in word affection compared with coarse-grained labels in binary polarity only (as done in LBSA and VADER).

5.5. Case study

To provide more insights to what can be learned by AFNN, this subsection presents a case study to qualitatively compare BiLSTM+AF and BiLSTM+LBSA, the current SOTA affective-driven method. Although both methods have attention mechanism to highlight sentiment lexicon conveying commonly agreed affective information, BiLSTM+AF models the interactions between automatic features with fine-grained affective annotations, whereas BiLSTM+LBSA employs positive and negative lexicon to set the attention weights.

⁷There is only one document under Scarce condition in IMDB due to the long comments given in this dataset, so the bar shows 0 or 1 in accuracy.

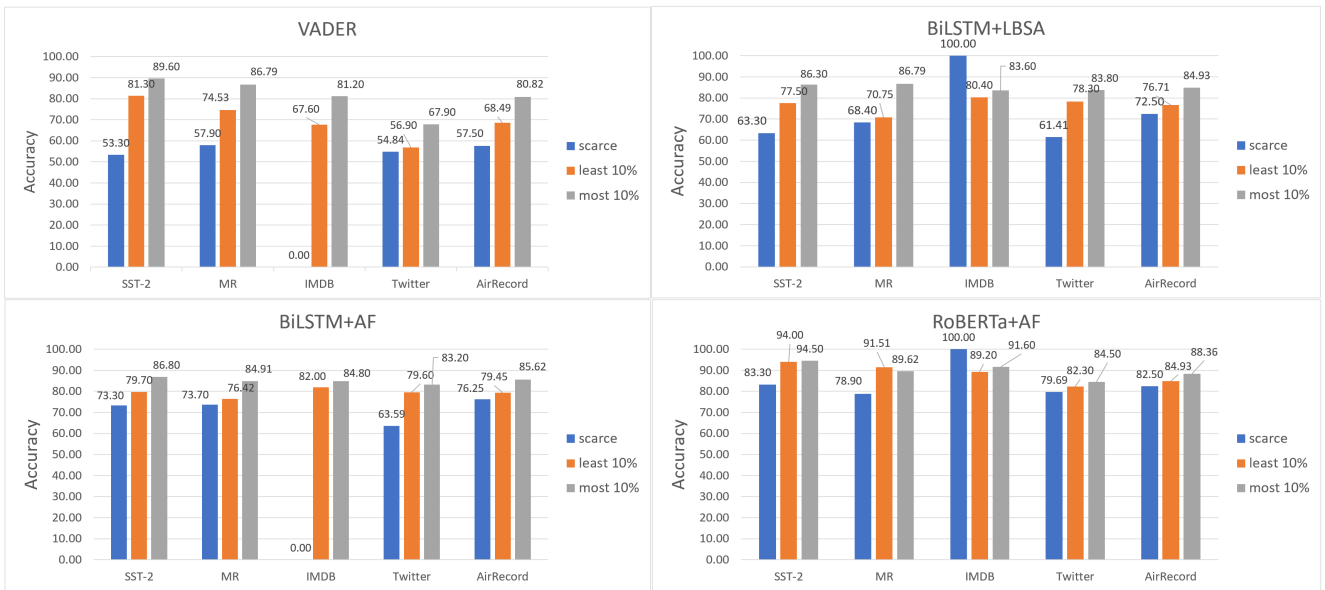


Figure 8: Accuracy over test samples with varying sentiment lexicon rates. The barplots from left to right shows the results for scarce, least 10%, and most 10% sentiment lexicon.

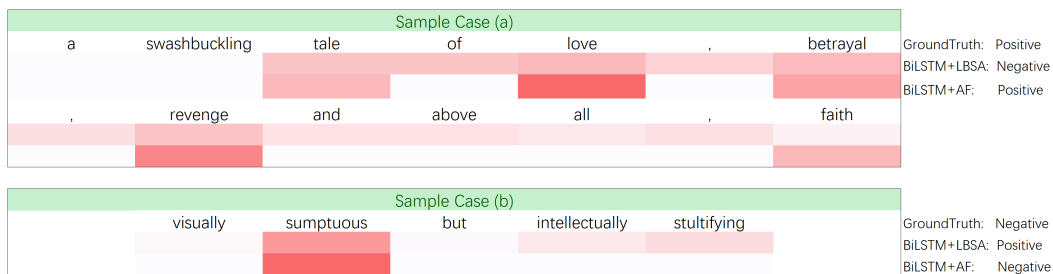


Figure 9: Heatmaps showing the attention weights from BiLSTM+LBSA (upper bar) and BiLSTM+AF (lower bar) for the cases in Figure 1. Darker reds indicate higher weights. BiLSTM+AF make correct predictions in both cases while BiLSTM+LBSA make mistakes owing to the heavy reliance on the sentiment lexicon.

We take the two examples in Figure 1 for the qualitative analyses. Recall that they exhibit opposite polarity and will help us understand what are learned for positive and negative sentiment. Figure 9 shows the heatmap visualizing the attention weight distributions.

In case (a), there are five affective terms from the lexicon — ‘tale’, ‘love’, ‘betrayal’, ‘revenge’, and ‘faith’, which are mostly emphasized with higher weights for both BiLSTM+LBSA and BiLSTM+AF. It is hence challenging to make sense of the overall sentiment given large proportion of affective words. This is because the models may use the explicit indicators (from human annotations) for prediction, although some labels maybe helpful while others misleading. BiLSTM+LBSA largely rely on the lexicon to assign attention weights, whereas the occurrences of negative words (‘betrayal’ and ‘revenge’) may hinder its capability to capture the global senses for correct predictions. On the contrary, BiLSTM+AF can gain affective knowledge from multiple dimensions beyond polarity, which therefore weigh ‘love’ over ‘betrayal’ and ‘revenge’, and successfully predict the

positive sentiment for the comment.

Case (b) demonstrates another trap possibly resulted from sentiment lexicon. In this short comment, ‘sumptuous’ is the only annotated lexicon available to the models, which is unsurprisingly assigned higher weights by both BiLSTM+AF and BiLSTM+LBSA. However, the uncommon word ‘stultifying’, though absent in the lexicon, plays an crucial role to indicate the overall sentiment polarity. Although both attentions fail to signal its importance (BiLSTM+LBSA seems to allocate some weights to it but very limited), our model correct the mistake at the hybrid learning stage with the help of automatic features. This proves potential benefits of coupling automatic features and handcrafted affective knowledge simultaneously, which may fix such errors.

6. Conclusion

This article has presented a novel AFNN for sentiment analysis via coupling manual annotations of word-level affective attributes and automatic semantic features learned by

neural networks. We argue that the fusion of regular representation learning and affective learning can benefit neural network to gain awareness of semantics and emotions, which are essential affective clues to identify the writers' sentiment from texts.

On the basis of the above hypothesis, we combine regular representation and affective driven learning networks to obtain both contextual semantics and affective awareness. We adopt the sentiment knowledge of EPA and VAD as fine-grained prior affective knowledge, which can be easily integrated into neural networks with minimal modification. Performance evaluations on various popular methods based on LSTM and SOTA pre-trained language models across five benchmark datasets indicate that AFNN can achieve 1.4% accuracy gain on average. Furthermore, the analysis of affective attributes demonstrates the salience of Evaluation and Valence for sentiment prediction, while integrating more attributes can further boost the overall performance. Next, the investigation on model sensitivity to affective lexicon rate shows that fusion mechanism can mitigate the over dependency problem to sentiment lexicons observed in many existing affective-driven models. Finally, the results from case studies show our superiority to leverage richer affective knowledge and automatic features to predict the overall sentiment.

Despite of the promising results, this work may still present some limitations in terms of 1) the diversity of the benchmark datasets, 2) the size and word coverage of the sentiment lexicons, and 3) the availability of more sentiment lexicons. Future work will continue to evaluate the proposed method on more corpora that text sources with richer annotations and larger scales. Besides, efforts will also be made for adapting our idea to the few shot learning settings. For example, we may consider to expand the affective knowledge lexicon with automatic annotation mechanisms, which will further mitigate the reliance on sentiment lexicon.

Acknowledgements

We thank the editors and all anonymous reviewers for the insightful suggestions on various aspects of this work. This work is supported by the research grants from Hong Kong Polytechnic University (PolyU RTVU, 1-BE2W, and 1-ZVRH), GRF grant (CERG PolyU 15211/14E, PolyU 152-006/16E), the Hong Kong Polytechnic University Postdoctoral Fellowships Scheme Projects #YW4H, NSFC Young Scientists Fund (62006203), and CCF-Tencent Rhino-Bird Young Faculty Open Research Fund.

References

- [1] Akhtar, M.S., Gupta, D., Ekbal, A., Bhattacharyya, P., 2017. Feature selection and ensemble construction: A two-step method for aspect based sentiment analysis. *Knowledge-Based Systems* 125, 116–135.
- [2] Akhtar, S., Ghosal, D., Ekbal, A., Bhattacharyya, P., Kurohashi, S., 2019. All-in-one: Emotion, sentiment and intensity prediction using a multi-task ensemble framework. *IEEE Transactions on Affective Computing*.
- [3] Alfrjani, R., Osman, T., Cosma, G., 2019. A hybrid semantic knowledgebase-machine learning approach for opinion mining. *Data & Knowledge Engineering* 121, 88–108.
- [4] Ali, F., Kwak, D., Khan, P., El-Sappagh, S., Ali, A., Ullah, S., Kim, K.H., Kwak, K.S., 2019. Transportation sentiment analysis using word embedding and ontology-based topic modeling. *Knowledge-Based Systems* 174, 27–42.
- [5] Andreevskaia, A., Bergler, S., 2008. When specialists and generalists work together: Overcoming domain dependence in sentiment tagging. *Proceedings of ACL-08: HLT*, 290–298.
- [6] Basiri, M.E., Nemati, S., Abdar, M., Cambria, E., Acharya, U.R., 2020. Abcdm: An attention-based bidirectional cnn-rnn deep model for sentiment analysis. *Future Generation Computer Systems* 115, 279–294.
- [7] Cambria, E., 2016. Affective computing and sentiment analysis. *IEEE intelligent systems* 31, 102–107.
- [8] Cambria, E., Das, D., Bandyopadhyay, S., Feraco, A., 2017. Affective computing and sentiment analysis, in: *A practical guide to sentiment analysis*. Springer, pp. 1–10.
- [9] Chen, H., Sun, M., Tu, C., Lin, Y., Liu, Z., 2016. Neural sentiment classification with user and product attention, in: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Austin, Texas. pp. 1650–1659.
- [10] Devlin, J., Chang, M., Lee, K., Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: *Proceedings of NAACL*, pp. 4171–4186.
- [11] Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., Gao, J., Zhou, M., Hon, H.W., 2019. Unified language model pre-training for natural language understanding and generation, in: *Advances in Neural Information Processing Systems*, pp. 13063–13075.
- [12] Farhadloo, M., Rolland, E., 2016. Fundamentals of sentiment analysis and its applications, in: *Sentiment Analysis and Ontology Engineering*. Springer, pp. 1–24.
- [13] Graves, A., Schmidhuber, J., 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks* 18, 602–610.
- [14] He, R., Lee, W.S., Ng, H.T., Dahlmeier, D., 2018. Effective attention modeling for aspect-level sentiment classification, in: *Proceedings of the 27th international conference on computational linguistics*, pp. 1121–1131.
- [15] Heise, D.R., 2010. *Surveying cultures: Discovering shared conceptions and sentiments*. John Wiley & Sons.
- [16] Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. *Neural computation* 9, 1735–1780.
- [17] Hogenboom, A., Heerschop, B., Frasinca, F., Kaymak, U., de Jong, F., 2014. Multi-lingual support for lexicon-based sentiment analysis guided by semantics. *Decision support systems* 62, 43–53.
- [18] Hovy, E.H., 2015. What are sentiment, affect, and emotion? applying the methodology of michael zock to sentiment analysis, in: *Language production, cognition, and the Lexicon*. Springer, pp. 13–24.
- [19] Hu, G., Lu, G., Zhao, Y., 2021. Fss-gcn: A graph convolutional networks with fusion of semantic and structure for emotion cause analysis. *Knowledge-Based Systems* 212, 106584.
- [20] Hu, X., Tang, J., Gao, H., Liu, H., 2013. Unsupervised sentiment analysis with emotional signals, in: *Proceedings of the 22nd international conference on World Wide Web*, pp. 607–618.
- [21] Huang, M., Xie, H., Rao, Y., Liu, Y., Poon, L.K., Wang, F.L., 2020. Lexicon-based sentiment convolutional neural networks for online review analysis. *IEEE Transactions on Affective Computing*.
- [22] Hutto, C., Gilbert, E., 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text, in: *Proceedings of the International AAAI Conference on Web and Social Media*, pp. 216–225.
- [23] Irsay, O., Cardie, C., 2014. Opinion mining with deep recurrent neural networks., in: *EMNLP*, pp. 720–728.
- [24] Ke, P., Ji, H., Liu, S., Zhu, X., Huang, M., 2020. Sentilare: Linguistic knowledge enhanced language representation for sentiment analysis,

- in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 6975–6988.
- [25] Khoo, C.S., Johnkhan, S.B., 2018. Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *Journal of Information Science* 44, 491–511.
- [26] Kim, Y., 2014. Convolutional neural networks for sentence classification, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar. pp. 1746–1751.
- [27] Liu, B., 2020. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge university press.
- [28] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- [29] Long, Y., Qin, L., Xiang, R., Li, M., Huang, C.R., 2017. A cognition based attention model for sentiment analysis, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 473–482.
- [30] Long, Y., Xiang, R., Lu, Q., Huang, C.R., Li, M., 2019. Improving attention model based on cognition grounded data for sentiment analysis. *IEEE Transactions on Affective Computing*, 1–14.
- [31] Loria, S., 2018. *textblob* documentation. Release 0.15.2.
- [32] Ma, Y., Peng, H., Khan, T., Cambria, E., Hussain, A., 2018. Sentic lstm: a hybrid network for targeted aspect-based sentiment analysis. *Cognitive Computation* 10, 639–650.
- [33] Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C., 2011. Learning word vectors for sentiment analysis, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, Association for Computational Linguistics. pp. 142–150.
- [34] Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S., McClosky, D., 2014. The stanford corenlp natural language processing toolkit., in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (System Demonstrations), pp. 55–60.
- [35] Melville, P., Gryc, W., Lawrence, R.D., 2009. Sentiment analysis of blogs by combining lexical knowledge with text classification, in: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM. pp. 1275–1284.
- [36] Meškeli, D., Frasincar, F., 2020. Aldonar: A hybrid solution for sentence-level aspect-based sentiment analysis using a lexicalized domain ontology and a regularized neural attention model. *Information Processing & Management* 57, 102211.
- [37] Mingyu, W., Ahrens, K., Chersoni, E., Jiang, M., Su, Q., Xiang, R., Huang, C.R., 2020. Using conceptual norms for metaphor detection, in: Proceedings of the Second Workshop on Figurative Language Processing, pp. 104–109.
- [38] Mohammad, S., 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 174–184.
- [39] Nazir, A., Rao, Y., Wu, L., Sun, L., 2020. Issues and challenges of aspect-based sentiment analysis: A comprehensive survey. *IEEE Transactions on Affective Computing*.
- [40] Osgood, C.E., Suci, G.J., Tannenbaum, P.H., 1957. *The measurement of meaning*. 47, University of Illinois press.
- [41] Pang, B., Lee, L., 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales, in: Proceedings of the 43rd annual meeting on association for computational linguistics, Association for Computational Linguistics. pp. 115–124.
- [42] Pennington, J., Socher, R., Manning, C.D., 2014. GloVe: Global Vectors for Word Representation, in: Proceedings of EMNLP, pp. 1532–1543.
- [43] Qian, Q., Huang, M., Lei, J., Zhu, X., 2017. Linguistically regularized lstm for sentiment classification, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1679–1689.
- [44] Rao, Y., Li, Q., Mao, X., Wenyan, L., 2014. Sentiment topic models for social emotion mining. *Information Sciences* 266, 90–100.
- [45] Russell, J.A., 2003. Core affect and the psychological construction of emotion. *Psychological review* 110, 145.
- [46] Smith-Lovin, L., Heise, D.R., 1988. *Analyzing Social Interaction: Advances in Affect Control Theory*. volume 13. Taylor & Francis.
- [47] Socher, R., Pennington, J., Huang, E.H., Ng, A.Y., Manning, C.D., 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics. pp. 151–161.
- [48] Socher, R., Perelygin, A., Wu, J.Y., Chuang, J., Manning, C.D., Ng, A.Y., Potts, C., 2013. Recursive deep models for semantic compositionality over a sentiment treebank, in: Proceedings of the conference on empirical methods in natural language processing (EMNLP), Citeseer. p. 1642.
- [49] Tang, D., Qin, B., Liu, T., 2015a. Document modeling with gated recurrent neural network for sentiment classification, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1422–1432.
- [50] Tang, D., Qin, B., Liu, T., 2015b. Learning Semantic Representations of Users and Products for Document Level Sentiment Classification, in: Proceedings of ACL-IJCNLP, pp. 1014–1023.
- [51] Tang, D., Wei, F., Qin, B., Zhou, M., Liu, T., 2014a. Building large-scale twitter-specific sentiment lexicon: A representation learning approach, in: Proceedings of coling 2014, the 25th international conference on computational linguistics: Technical papers, pp. 172–182.
- [52] Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., Qin, B., 2014b. Learning sentiment-specific word embedding for twitter sentiment classification., in: *ACL (1)*, pp. 1555–1565.
- [53] Tang, Y.j., Chen, H.H., 2012. Mining sentiment words from microblogs for predicting writer-reader emotion transition., in: *LREC*, pp. 1226–1229.
- [54] Teng, Z., Vo, D.T., Zhang, Y., 2016. Context-sensitive lexicon features for neural sentiment analysis., in: *EMNLP*, pp. 1629–1638.
- [55] Wan, M., Xing, B., 2020. Modality enriched neural network for metaphor detection, in: Proceedings of the 28th International Conference on Computational Linguistics, pp. 3036–3042.
- [56] Wang, Y., Huang, M., Zhu, X., Zhao, L., 2016. Attention-based lstm for aspect-level sentiment classification, in: Proceedings of the 2016 conference on empirical methods in natural language processing, pp. 606–615.
- [57] Wilson, T., Wiebe, J., Hoffmann, P., 2005. Recognizing contextual polarity in phrase-level sentiment analysis, in: Proceedings of the conference on human language technology and empirical methods in natural language processing, Association for Computational Linguistics. pp. 347–354.
- [58] Xiang, R., Long, Y., Wan, M., Gu, J., Lu, Q., Huang, C.R., 2020. Affection driven neural networks for sentiment analysis, in: Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), European Language Resources Association. pp. 112–119.
- [59] Yadollahi, A., Shahraki, A.G., Zaiane, O.R., 2017. Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys (CSUR)* 50, 1–33.
- [60] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V., 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems* 32, 5753–5763.
- [61] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E., 2016. Hierarchical attention networks for document classification, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1480–1489.
- [62] Zhang, Y., Wang, J., Zhang, X., 2018. Ynu-hpec at semeval-2018 task 1: Bilstm with attention based sentiment analysis for affect in tweets, in: Proceedings of The 12th International Workshop on Semantic Evaluation, pp. 273–278.

- [63] Zou, Y., Gui, T., Zhang, Q., Huang, X., 2018. A lexicon-based supervised attention model for neural sentiment analysis, in: Proceedings of the 27th International Conference on Computational Linguistics, pp. 868–877.