Affection Driven Neural Networks for Sentiment Analysis

Rong Xiang[1], Yunfei Long[34], Mingyu Wan[25], Jinghang Gu[2], Qin Lu[1], Chu-Ren Huang[2]

The Hong Kong Polytechnic University[12], University of Nottingham[3], Peking University[4], University of Essex[5]
Department of Computing, 11 Yuk Choi Road, Hung Hom, Hong Kong (China)[1]
Chinese and Bilingual Studies, 11 Yuk Choi Road, Hung Hom, Hong Kong (China)[2]
NIHR Nottingham Biomedical Research Centre, Nottingham (UK) [3]
Engineering, University of Essex, Colchester (UK)[4]
School of Foreign Languages, 5 Yiheyuan Road, Haidian, Beijing (China)[5]
School of Computer Science and Electronic xiangrong0302@gmail.com[1], qin.lu@comp.polyu.edu.hk[1],
{mingyu.wan,jinghgu,churen.huang}@polyu.edu.hk[2], yunfei.long@nottingham.ac.uk[3]

Abstract

Deep neural network models have played a critical role in sentiment analysis with promising results in the recent decade. One of the essential challenges, however, is how external sentiment knowledge can be effectively utilized. In this work, we propose a novel affection-driven approach to incorporating affective knowledge into neural network models. The affective knowledge is obtained in the form of a lexicon under the Affect Control Theory (ACT), which is represented by vectors of three-dimensional attributes in Evaluation, Potency, and Activity (EPA). The EPA vectors are mapped to an affective influence value and then integrated into Long Short-term Memory (LSTM) models to highlight affective terms. Experimental results show a consistent improvement of our approach over conventional LSTM models by 1.0% to 1.5% in accuracy on three large benchmark datasets. Evaluations across a variety of algorithms have also proven the effectiveness of leveraging affective terms for deep model enhancement.

Keywords: Sentiment analysis, Affective knowledge, Attention mechanism

1. Introduction

Texts are collections of information that may encode emotions and deliver impacts to information receivers. Recognizing the underlying emotions encoded in a text is essential to understanding the key information it conveys. On the other hand, emotion-embedded text can also provide rich sentiment resources for relevant natural language processing tasks. As such, sentiment analysis (SA) has gained increasing interest among researchers who are keen on the investigation of natural language processing techniques as well as emotion theories to identify sentiment expressions in a natural language context. Typical SA studies analyze subjective documents from the author's perspective using high-frequency word representations and mapping the text (e.g., sentence or document) to categorical labels, e.g., sentiment polarity, with either a discrete label or a real number in a continuum.

Recently, the rising use of neural network models has further elevated the performance of SA without involving laborious feature engineering. Typical neural network models such as Convolutional Neural Network (CNN) (Kim, 2014), Recursive auto-encoders (Socher et al., 2013), Long-Short Term Memory (LSTM) (Tang et al., 2015a) have shown promising results in a variety of sentiment analysis tasks. In spite of this, neural network models still face two main problems. First, neural network approaches lack direct mechanisms to highlight important components in a text. Second, external resources such as linguistic knowledge, cognition grounded data, and affective lexicons, are not fully employed in neural models. To tackle the first problem, cognition-based attention models have been adopted for sentiment classification using text-embedded information such as users, products, and local context (Tang

et al., 2015b; Yang et al., 2016; Chen et al., 2016; Long et al., 2019). For the second problem, Qian et al. (2016) proposed to add linguistic resources to deep learning models for further improvement. Yet, recent method of integration of additional lexical information are limited to matrix manipulation in attention layer due to the incompatibility of such representations with embedding ones, making it quite inefficient. In this paper, we attempt to address this problem by incorporating an affective lexicon as numerical influence values into affective neural network models through the framework of the Affect Control Theory (ACT).

ACT is a social psychological theory pertaining to social interactions (Smith-Lovin and Heise, 1988). It is based on the assumption that people tend to maintain culturally shared perceptions of identities and behaviors in transient impressions during observation and participation of social events (Joseph, 2016). In other words, social perceptions, actions, and emotional experiences are governed by a psychological intention to minimize deflections between fundamental sentiments and transient impressions that are inherited from the dynamic behaviors of such interactions. To capture such information, an event in ACT is modeled as a triplet: {actor, behavior, object}. In other words, culturally shared "fundamental" sentiments about each of these elements are measured in three dimensions: **Evaluation**, **Potency**, and **Activity**, commonly denoted as (**EPA**).

In the ACT theory, emotions are functions of the differences between fundamental sentiments and transient impressions. The core idea is that each of the entities {actor, behavior, object} in an event has a fundamental emotion (EPA value) that is shared among members of the same culture or community. All the entities in the event as a group generate a transient impression or feeling that might be different from the fundamental sentiment. Previous research (Osgood et

al., 1964) used EPA profiles of concepts to measure semantic differential, a survey technique to obtain respondent rates in terms of affective scales (e.g., {*good, nice*}, {*bad, awful*} for E; {*weak, little*}, {*strong, big*} for P; {*calm, passive*}, {*exciting, active*} for A). Existing datasets with average EPA ratings are usually small-sized, such as the dataset provided by Heise (2010) which compiled a few thousands of words from participants of sufficient cultural knowledge. As an illustration of the data form, the culturally shared EPA for the word "mother" is [2.74, 2.04, 0.67], which corresponds to {*quite good*}, {*quite powerful*}, and {*slightly active*}.

ACT has demonstrated obvious advantages for sentiment analysis. This is because the model is more cognitively sound and the representation is more comprehensive. The use of ACT is especially effective in long documents, such as review texts composed by a lot of descriptive events or factual reports(Heise, 2010). Being empirically driven, EPA space enables the universal representation of individuals' sentiment, which can reflect a real evaluation of human participants. More importantly, the interaction between terms in ACT complies with the linguistic principle of the compositional semantic model that the meaning of a sentence is a function of its words (Frege, 1948).

In line with the above framework, we propose an affection driven neural network method for sentiment classification by incorporating an ACT lexicon as additional affective knowledge into deep learning models. Instead of transforming sentiment into a feature matrix, we apply EPA values into numeric weights directly so that it is more efficient. Our approach can be generally applied to a wide range of current deep learning algorithms. Among different deep learning algorithms, we choose to demonstrate our method using LSTM as it is quite suited for NLP with proven performance. A series of LSTM models are implemented without using dependency parsing or phrase-level annotation. We identify affective terms with EPA values and transform their corresponding EPA vectors into a feature matrix. Single EPA values are then computed and integrated into deep learning models by a linear concatenation. Evaluations are conducted in three sentiment analysis datasets to verify the effectiveness of the affection-driven network.

The rest of this paper is organized as follows. Section 2 introduces related works in Affect Control Theory, deep learning classifiers and some use of affective knowledge. Section 3 describes detailed design of our proposed method. Performance evaluation and analysis are presented in Section 4. Section 5 concludes the paper with some possible directions for future works.

2. Related Work

Related work is reviewed through three sub-sections: affective lexicons under the framework of affective control theory, general studies in neural network sentiment analysis, and specific research of using lexicon knowledge for sentiment analysis.

2.1. Affective Lexicons under ACT

Affective lexicons with EPA values are available for many languages, yet all are small-sized. Previous works suggest that the inter-cultural agreement on EPA meanings of social concepts is generally high even across different subgroups of society. Cultural-average EPA ratings from a few dozen survey participants have proven to be extremely stable over extended periods (Heise, 2010). These findings shed some light on the societal conflicts by competing for political ideologies. It is also proved that the number of contested concepts is small relative to the stable and consensual semantic structures that form the basis of our social interactions and shared cultural understanding (Heise, 2010).

To date, the most reliable EPA based affective lexicons are obtained by manual annotation. For example, the EPA lexicon provided by Heise (1987) are manually rated in the evaluation-potency-activity (EPA) dimensions. Although, there is no size indication, this EPA based lexicon is commonly used as a three dimensional affective resource. This professional annotated lexicon are regarded as a high-quality lexicon (Bainbridge et al., 1994) and it the main resource used in this work as the external affective resource. In 2010, a new release of this resource includes a collection of five thousand lexical items [1] (Heise, 2010).

2.2. Deep Neural Networks

In recent years, neural network methods have greatly improved the performance of sentiment analysis. Commonly used models include Convolutional Neural Networks (CNN) (Socher et al., 2011), Recursive Neural Network ReNN (Socher et al., 2013), and Recurrent Neural Networks (RNN) (Irsoy and Cardie, 2014). Long-Short Term Memory model (LSTM), well known for text understanding, is introduced by Tang et al. (2015a) who added a gated mechanism to keep long-term memory. Attention-based neural networks, mostly built from local context, are proposed to highlight semantically important words and sentences (Yang et al., 2016). Other methods build attention models using external knowledge, such as user/product information (Chen et al., 2016) and cognition grounded data (Long et al., 2019).

2.3. Use of Affective Knowledge

Previous studies in combining lexicon-based methods and machine learning approach generally diverge into two ways. The first approach uses two weighted classifiers and linearly integrates them into one system. Andreevskaia and Bergler (2008), for instance, present an ensemble system of two classifiers with precision-based weighting. This method obtained significant gains in both accuracy and recall over corpus-based classifiers and lexicon-based systems. The second approach incorporates lexicon knowledge into learning algorithms. To name a few, Hutto and Gilbert (2014) design a rule-based approach to indicate sentiment scores. Wilson et al. (2005) and Melville et al. (2009) use a general-purpose sentiment dictionary to improve linear classifier. Jovanoski et al. (2016) also prove that sentiment lexicon can contribute to logistic regression models. In neural network models, a remarkable

---

work on utilizing sentiment lexicons is done by Teng et al. (2016). They treat the sentiment score of a sentence as a weighted sum of prior sentiment scores of negation words and sentiment words. Qian et al. (2016) propose to apply linguistic regularization to sentiment classification with three linguistically motivated structured regularizers based on parse trees, topics, and hierarchical word clusters. Zou et al. (2018) adopt a mixed attention mechanism to further highlight the role of sentiment lexicon in the attention layer. Using sentiment polarity in a loss function is one way to employ attention mechanism. However, attention weights are normally obtained using local context information. The computational complexity of reweighing each word by attention requires matrix and softmax manipulation, which slows down the time for training and inference especially with long sentences.

3. Methodology

We proposes a novel affection driven method for neural sentiment classification. The affective lexicon with EPA values is used as the external affective knowledge which is integrated into neural networks for performance enhancement. The use of external knowledge reduces computation time and the cognition grounded three dimensional affective information using EPA is more comprehensive.

The method works as follows. The affective terms in a task dataset are first identified in order to collect their EPA vectors through a pre-processing step. Each identified affective term is then given a weight based on a linear transformation mapping the three dimensional EPA values into a single value with a corresponding affective polarity. The affective weight will grant the prior affective knowledge to the identified affective terms to enhance word representation as a coefficient. This set of affective coefficients are used to adjust the weights in neural network models. This work applies the affective coefficients to a number of LSTM models including the basic LSTM, LSTM with attention layer (LSTM-AT), Bi-direction LSTM (BiLSTM) and BiLSTM with attention layer (BiLSTM-AT) with the EPA weights. This mechanism is generally applicable to many neural network models.

3.1. EPA Weight Transformation

In this work, we use the affective lexicon with EPA values provided by Heise (2010) [2] For each term, the EPA values are measured separately in three separate dimensions as numerical values in the continuous space ranging from -4.50 to 4.50. The signum indicates the correlation, while the value implies the degree of relation. Figure 1 below shows the histograms of the three affective measures in Heise's work.

As the histogram suggests, none of the three measures E, P, and A shows a balanced distribution, and they are overall right-skewed. The evaluation component is the most evenly distributed amongst all. Notably, the Evaluation distribution has two peaks scattered at both the positive axis and
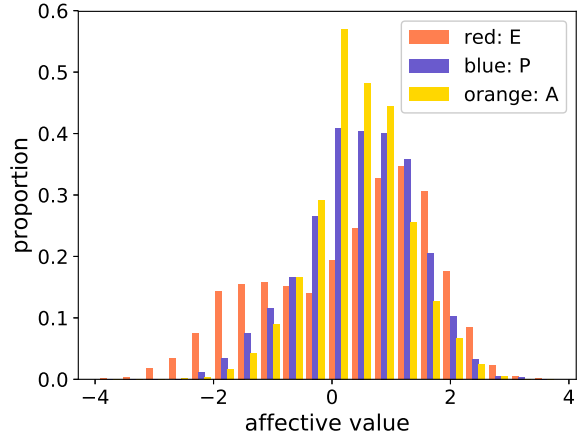

Figure 1: Histogram of EPA values

the negative axis, which is apparently different from Potency and Activity. Potency and Activity generally follow Gaussian distribution and the means are around 0.60. The majority of the affective values fall in the range of -2.20 to 3.20, yet their variances are largely different. Most Activity values are distributed near the mean, providing less significant evidence for affective expression.

When using EPA to identify the polarity of sentiment, the E, P, and A weights need to be projected to one value before integrating it into a deep learning model. As a result, the EPA values are transformed into one single weight W_{EPA} which is regarded as an affective influence value. Affine combination is used to constrain this value to stay in the range of [-4.50, 4.50], as formulated below:

$$W_{EPAcomb} = \alpha W_E + \beta W_P + \gamma W_A, \quad (1)$$

where

$$\alpha + \beta + \gamma = 1, \quad (2)$$

and α, β, γ are hyper parameters to indicate the significance of each component. For instance, we use [1, 0, 0] to indicate the exclusive use of Evaluation. To avoid the over-weighting problem for affective terms and at the same time to highlight the intensity information of EPA values, another linear transformation is defined below:

$$W_{EPA} = (1 + a|W_{EPAcomb}|). \quad (3)$$

W_{EPA} is a weight value, referred as the **affective influence value**. Equation 3 ensures that all terms in the EPA lexicon will have value over one. Terms in the target dataset which do not appear in the EPA lexicon will have the weight value of one. a is a non-negative parameter that can be tuned as the amplification of EPA values.

3.2. Affective Neural Network Framework

This section elaborates on the mechanism of our proposed affective neural networks. In other words, how we incorporate affective influence values into affective deep neural networks. Although neural network models such as LSTM with added attention layer has powerful learning capacity

---

[2]http://www.indiana.edu/~socpsy/public_files/EnglishWords_EPAs.xlsx which covers the most commonly-used five thousand manually annotated English sentiment words.

for sentences, they demonstrate no explicit ability in identifying lexical sentiment affiliation. Serving as prior affective knowledge, these affective influence values can be used in any deep learning model. Figure 2 shows the general framework of our proposed method to incorporate affective influence values into any neural network model that involves the learning of word representation. Simply put in a deep learning model, the learning of word representation is carried out in the word representation layer to obtain their representations $\hat{h}_i$. The affective influence values as representation of affective information W_{EPAi} is then incorporated with $\hat{h}_i$ before it goes into the pooling layer.
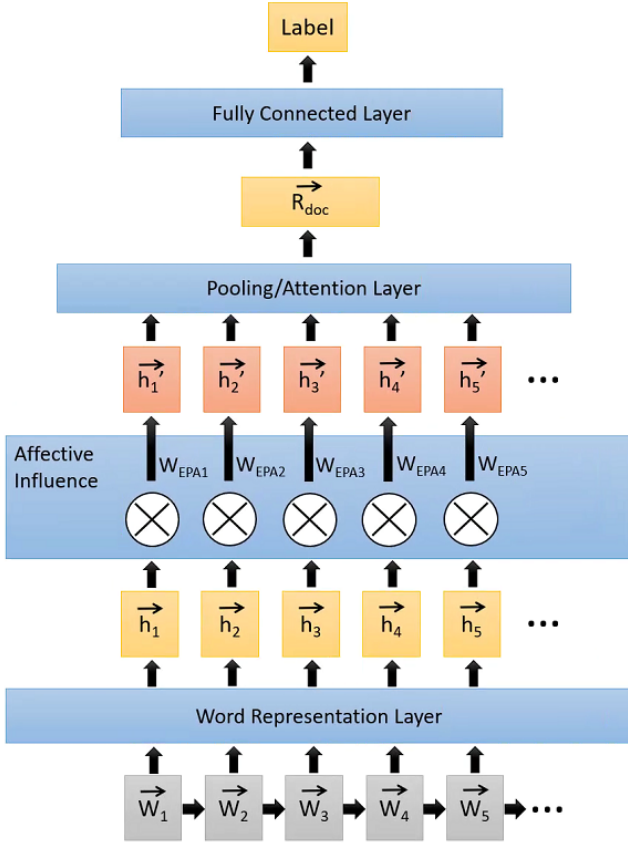

Figure 2: Framework of affective deep learning schema

Let D be a collection of n documents for sentiment classification. Each document d_i is an instance in D, ($i \in 1, 2, ..., n$). In sentiment analysis, the label can either be a binary value to simply indicate polarity or a numerical value to indicate both polarity and strength. Each document d_i is first tokenized into a word sequence. The representation vector of words, denoted as $\overrightarrow{w_i}$, is then obtained from a word embedding layer.

For LSTM-based algorithms, the word representation vectors h_i are updated in the recurrent layer. To incorporate affective knowledge, we use the product of W_{EPAi} with its corresponding word representation $\overrightarrow{h_i}$.

$$\overrightarrow{h_i}' = W_{EPAi} * \overrightarrow{h_i} \qquad (4)$$

This computation will be repeated during the entire process. As a result, all recognized affective terms are highlighted with certain intensity. The updated word representation $\overrightarrow{h_i}'$ can then be fed to the pooling layer or attention layer as usual, generating document representation $\overrightarrow{R_{doc}}$. Thus, $\overrightarrow{R_{doc}}$ accommodates both semantic information and affective prior knowledge for the classifier layer.

Using W_{EPA} as attention weight can significantly accelerate the training and inference speed compared to methods of using local context to get attention weights. This is because getting W_{EPAi} as a linear transformation only takes constant time so that it is not related to document size. Incorporating W_{EPAi} also takes a fixed time. However, for getting attention weights for n length documents, it requires matrix operation whose calculation required $O(n)$.

4. Performance Evaluation

Performance evaluation is conducted on three benchmark datasets[3], including a Twitter collection, an airline dataset, and an IMDB review. The baseline classifiers include Support Vector Machine (SVM), CNN, LSTM, and BiLSTM. Attention-based LSTM and BiLSTM are also implemented for further comparisons.

4.1. Datasets and Settings

The first benchmark dataset (Twitter) is collected from twitter and is publicly available for sentiment analysis[4]. The content is mainly about personal opinions on events, gossips, etc. The affective labels are defined as binary values to indicate positive and negative polarities.

The second benchmark dataset (AirRecord) consists of customer twitted messages from six major U.S. airlines[5]. It includes 14,640 messages collected in February of 2015 which were manually labeled with positive, negative, and neutral classes.

The third dataset (IMDB) is collected and provided by Maas et.al (2011), which contains user comments of paragraphs extracted from online IMDB film database[6]. Affective labels are binary values for positive and negative.

To utilize the affective lexicon, all the three datasets are pre-processed to identify affective terms in the affective lexicon. Table 1 shows some statistical data of the three datasets including the proportions of affective terms over the total number of words in the datasets.

Dataset name	Instance total	Average length	Affective terms %
Twitter	99,989	13.7	20.2%
AirRecord	14,640	17.8	18.1%
IMDB	25,000	259.5	21.1%

Table 1: Datasets Statistics

Sentences in both Twitter and AirRecord are relatively short whereas sentences in IMDB are ten times longer on average. In all three datasets, EPA sentiment terms account for

[3]The three datasets are all publicly available in Kaggle: https://www.kaggle.com/

[4]https://www.kaggle.com/c/twitter-sentiment-analysis2

[5]https://www.kaggle.com/crowdflower/twitter-airline-sentiment/home/

[6]https://www.imdb.com/

about one-fifth of all words used in the dataset. Among the three datasets, affective terms take the largest proportion in IMDB while the percentage in AirRecord is 3% lower.

Both Twitter and IMDB are binary classification tasks and their performance is measured by accuracy and RMSE (rooted mean square error). Since AirRecord is labeled by positive, negative, or neutral, the F1 score is also provided. We take pre-trained Glove vectors (Pennington et al., 2014) as word embedding for deep learning models. As variants of LSTM are widely used in text classification tasks, we evaluate our methods on both the basic LSTM model and the BiLSTM model. Two LSTM variants with attention mechanism are also included in the evaluation, denoted as LSTM-AT and BiLSTM-AT respectively. All models are tuned with the three datasets. For a fair comparison, the parameters are set as follows:

Embedding size=300 dim, *Optimizer=Adam*, *Learning Rate=5e-4*, *Dropout=0.1*, *Batch Size=32* and *Epoch=3*. *Convolutional Kernel Size* for CNN is set to 3.

4.2. Sentiment Analysis

Partitioned by stratified sampling, 90% instances of each dataset are used as the training data and the remaining 10% serves as the testing data. Experiments are conducted to evaluate the prediction performance of LSTM algorithms using our proposed method compared with baseline models listed below. The average result of three runs for each setting are reported. By default, the hyper-parameters α, β and γ in Formula 1 and 2 are set equivalently to 0.33, and a is experimentally set to 1.15 as the optimized setting. The name of a model augmented with EPA values is expanded with (EPA).

- **SVM** is the basic model that uses a sentence feature vector. We use the mean of word embedding to generate the sentence representation.

- **CNN** uses a convolution layer to capture features of adjacent words. The final sentiment label is classified with a perceptron.

- **LSTM** is a typical RNN architecture with a gated mechanism. LSTMs were developed to handle exploding and vanishing gradient problems when training traditional RNNs.

- **LSTM-AT** uses LSTM with attention mechanism to re-weight important words before the fully connected layer.

- **BiLSTM** learns bidirectional long-term dependencies between time steps of sequential data. These dependencies can be useful for learning from the complete time series.

- **BiLSTM-AT** is BiLSTM with attention mechanism included. It is designed to combine strengths of both BiLSTM and attention mechanism.

Table 2 shows the overall performance for all the models for all the three datasets. SVM performs the worst among all approaches. This is because using mean word embedding vectors cannot fully capture contextual information in sentences. As a deep learning model, CNN shows some improvement compared to SVM. Although introducing a convolution window can include some adjacent lexical information in context, it lacks complete sentential representations. Also, the fixed length of the convolution window may work incorrectly on semantic segments. Thus the model can be hampered by additional noise intrinsic in the model. LSTM-based models, on the other hand, can manage to track long-term dependency and partially solve the vanishing gradient problem. As shown in Table 2, LSTM-based models significantly outperform CNN, and BiLSTM outperforms LSTM. This can be accounted for by the richer reverse information in BiLSTM. The attention mechanism also outperforms the baseline because it can put more emphasis on semantically salient terms. Considering these two variants of LSTM with bi-direction and attention, the improvement of attention mechanism is generally larger than bidirectional learning. Even though the results of bidirectional approaches are slightly inferior to unidirectional models, BiLSTM-AT with EPA generally achieves the best performance.

Let us take a closer look at how the performance varies on three benchmark datasets across different LSTM variants. In Twitter and AirRecord, the increased accuracy by using EPA values attains to 1% on average, and the improvement is even larger in IMDB (1.5%), which contains longer paragraphs of movie reviews. In particular, the BiLSTM model achieves a top performance with accuracy increased by 2.5%. As for the attention mechanism, it can potentially identify implicit semantic information from the local context, which can be used to adjust the coefficient of the word representation. Results of the four LSTM-based methods, however, suggest that non-attention models with EPA show more significant improvement than attention-based models. It indicates that using affective lexicon can be more effective than using attention. All the results congruently suggest that affective terms with informative sentiment representation can effectively and consistently contribute to model enhancement across different datasets, including both short sentences and long paragraphs, of which the later is more significant. Therefore, highlighting the affective terms relevant to sentiment could further improve the attention mechanism.

4.3. Ablation Analysis on EPA

Note that in Formula 1, there are three hyper parameters α, β, and γ. To see how different values of these hyper parameters affect the performance, we conducted the second set of experiments with different settings as follows:

- **EPA** follows the setting [0.33, 0.33, 0.33]

- **E** [1, 0, 0] is used for Evaluation only.

- **P** [0, 1, 0] is used for Potency only.

- **A** [0, 0, 1] is used for Activity only.

	Twitter		AirRecord			IMDB	
	ACC	RMSE	ACC	RMSE	F1	ACC	RMSE
SVM	0.548	NA	0.636	NA	0.48	0.498	NA
CNN	0.696	0.454	0.732	0.567	0.688	0.621	1.307
LSTM	0.747	0.327	0.805	0.428	0.721	0.803	0.405
LSTM(EPA)	0.759	0.318	**0.823**	0.410	**0.747**	0.818	0.398
LSTM-AT	0.754	0.321	0.811	0.425	0.733	0.807	0.395
LSTM-AT(EPA)	<u>0.761</u>	0.309	<u>0.820</u>	0.413	0.739	<u>0.819</u>	<u>0.392</u>
BiLSTM	0.756	0.314	0.807	0.418	0.739	0.797	0.404
BiLSTM(EPA)	**0.766**	<u>0.302</u>	0.817	0.411	0.736	**0.822**	<u>0.392</u>
BiLSTM-AT	0.759	0.317	0.813	**0.403**	0.743	0.805	<u>0.392</u>
BiLSTM-AT(EPA)	**0.766**	**0.301**	<u>0.820</u>	<u>0.409</u>	<u>0.745</u>	**0.822**	**0.387**

Table 2: Performance of Sentiment Analysis; global best is bolded; second-best is underlined

	Twitter		AirRecord			IMDB	
	ACC	RMSE	ACC	RMSE	F1	ACC	RMSE
LSTM	0.747	0.327	0.805	0.428	0.721	0.803	0.405
LSTM(E)	0.756	0.322	0.822	<u>**0.405**</u>	0.735	<u>0.819</u>	<u>0.394</u>
LSTM(P)	<u>**0.761**</u>	<u>**0.317**</u>	0.812	0.412	0.734	<u>**0.824**</u>	<u>**0.387**</u>
LSTM(A)	0.758	0.319	<u>**0.825**</u>	0.413	<u>0.739</u>	0.816	0.399
LSTM(EPA)	<u>0.759</u>	<u>0.318</u>	<u>0.823</u>	<u>0.410</u>	**0.747**	0.818	0.398
LSTM-AT	0.754	0.321	0.811	0.425	0.733	0.807	0.395
LSTM-AT(E)	0.758	0.317	**0.828**	<u>**0.402**</u>	**0.745**	**0.820**	**0.383**
LSTM-AT(P)	<u>0.759</u>	**0.307**	0.816	0.409	0.739	0.815	<u>0.387</u>
LSTM-AT(A)	<u>0.759</u>	0.314	<u>0.823</u>	<u>0.405</u>	<u>0.743</u>	0.817	0.389
LSTM-AT(EPA)	**0.761**	<u>0.309</u>	0.82	0.413	0.739	<u>0.819</u>	0.392
BiLSTM	0.756	0.314	0.807	0.418	0.739	0.797	0.404
BiLSTM(E)	0.760	0.313	<u>0.817</u>	**0.408**	**0.742**	0.817	0.399
BiLSTM(P)	<u>0.765</u>	**0.302**	**0.822**	<u>0.409</u>	<u>0.741</u>	0.815	0.398
BiLSTM(A)	0.764	<u>0.303</u>	0.814	0.413	0.733	<u>0.819</u>	<u>0.393</u>
BiLSTM(EPA)	**0.766**	<u>**0.302**</u>	0.817	0.411	0.736	**0.822**	<u>**0.392**</u>
BiLSTM-AT	0.759	0.317	0.813	0.403	<u>0.743</u>	0.805	0.392
BiLSTM-AT(E)	0.761	0.311	0.813	<u>0.398</u>	0.739	<u>0.822</u>	**0.381**
BiLSTM-AT(P)	0.762	<u>0.302</u>	<u>0.816</u>	0.405	0.742	**0.826**	<u>0.383</u>
BiLSTM-AT(A)	<u>0.764</u>	0.305	0.814	**0.396**	0.737	0.821	0.389
BiLSTM-AT(EPA)	**0.766**	**0.301**	**0.820**	0.409	**0.745**	<u>0.822</u>	0.387

Table 3: Performance of EPA components; overall best result is bolded; group best is marked bold with underline; second-best of each group is underlined

In ACT, the three components {E, P, A} are used as a group of information to characterize an affective related event. This section probes into the individual role of each attribute to the performance improvement. For each LSTM variant, Table 3 shows the evaluation of above four settings given in Formula 1 and 2 of the hyper-parameters α, β and γ.

As shown in 3, all affection-driven methods outperform their baseline counterparts. However, the performance discrepancy among different EPA components varies for different datasets. In Twitter, the accuracy of every component is very close (e.g., minor performance gaps in the range of 0.0%-0.2%) and EPA as a whole is the best (around 76%). In AirRecord, all four settings show similar performance (around 82%) although LSTM-AT with Evaluation gives the best performance (82.8%). For IMDB, the performance discrepancy is larger (in the range of 0.2%-0.8%). Among the three components, Evaluation shows slightly stronger effectiveness and Activity contributes the most on average. The overall performance suggests no superior component,

whereas, for different datasets, the proportion of E, P, and A can be fine-tuned to achieve finer improvement. One significant observation is EPA together serves as the best representation which indicates the orthogonality of each component could be supplementary to each other. Thus, we use EPA with equal weights for our model comparison in Section 4.2..

4.4. Case Study

This section studies the impact of affective lexicon to sentiment classification in comparison to the attention mechanism, which is known as a fine-tuned mechanism for word representation in the learning process. Affective lexicon contains external knowledge with words of commonly agreed affective values that without contextual information. In contrast, the attention mechanism aims to capture contextual information to optimize the weight of specific words. To make a clear comparison, we randomly select two sentences of distinct sentiment polarity, one positive

and the other negative, and show their weight distribution as heat graphs in Figures 3 and 4, respectively.
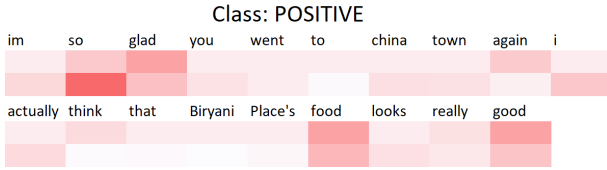

Figure 3: Sample Case 1. Red refers to a larger weight. White refers to a smaller weight

Figure 3 is the heat graph of sample case 1 with a positive polarity: "*I am so glad you went to China town again.I am actually think that Biryani Place's food looks really good*". The upper bar under the sentence indicates the recognized terms in an affective influence value sequence. Words not in the affective lexicon are displayed in white. The lower bar shows the heat map of the weighted word by attention. As Figure 3 clearly shows, affective terms account for a very small proportion. The words of red color in the EPA lexicon, i.e., '*glad*', '*again*', '*food*' and '*good*' are emphasized with higher weights. As the influence values of these terms are increased, the remaining words are less weighted by proportion. The attention weights are largely consistent with affective influence values, yet their intensity values are not as significant as the affective ones. Note that the attention mechanism puts more weight on the word '*so*', and less but still considerable weight on 'im', 'china town' and 'i'. These words may be semantically more important in the sentence. But, they are not necessarily related to sentiment expression.
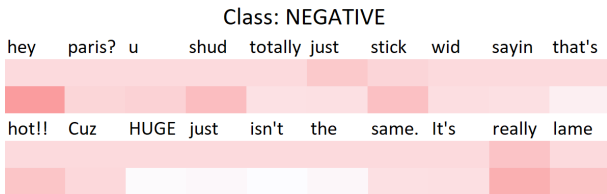

Figure 4: Sample Case 2. Red refers to a larger weight. White refers to a smaller weight

Figure 4 showcase the sample of a negative polarity: "*Hey, Paris? Ushud totally just stick wid sayin that's hot! Cuz HUGE just isn't the same. Its's really lame.*". Two adverbs '*just*' and '*really*' and one adjective '*hot*' are recognized in this sentence. However, the evidence '*lame*' is not in affective lexicon. Failing to identify this negative adjective made the affective knowledge base fail to update affective polarity. On the contrast, the attention mechanism does not require prior knowledge, and it can still identify '*lame*' as strong evidence.

In summary, affective knowledge in the form of a lexicon provides salient and reliable lexicon-level evidence for sentiment analysis. On the other hand, limited lexical coverage can lead to negative impact on updating word representation. Attention mechanism can be used as a self-adaptive method to highlight some important words if external knowledge is not available. Contextual patterns can

serve as supplementary information to word representation. In spite of the fact that words with high attention weight may be semantically more meaningful, such words may not directly related to sentiment. To aggregate the strengths of both methods, future attempts can be targeted at models with the incorporation of both mechanisms.

5. Conclusion and Future Work

This paper presents an affection-driven neural network by loaning an external lexicon that contains explicit sentiment prior knowledge framed under the Affect Control Theory. The external knowledge is cognition grounded and comprehensive. The method used can be easily integrated into deep learning models with only minimal computational cost. Performance evaluations on various LSTM based methods have congruently validated the hypothesis that Affective words with attributes of Evaluation, Potency, and Activity are more effective for sentiment analysis than other deep learning models, including attention-based LSTMs. The ablation introspection to the respective role of the three components in the EPA model suggests an equal contribution of them to model enhancement. To this end, Affine transformation of E, P, A is set with equal weight which in the end achieves the best performance in general.

Given the limitation of the size and coverage of the lexicon in this work, future efforts can be done in three directions. The first one is to further evaluate the performance of affection-driven neural networks on corpora of richer, larger and general text sources. The second one is to develop automatic annotation tools to scale up the EPA knowledge lexicon with a wider lexical coverage so as to further attest its effectiveness for sentiment analysis. The third one is to improve the methods by considering different EPA transformation functions for mapping into affect influence values.

6. Acknowledgements

7. Bibliographical References

Andreevskaia, A. and Bergler, S. (2008). When specialists and generalists work together: Overcoming domain dependence in sentiment tagging. *Proceedings of ACL-08: HLT*, pages 290–298.

Bainbridge, W. S., Brent, E. E., Carley, K. M., Heise, D. R., Macy, M. W., Markovsky, B., and Skvoretz, J. (1994). Artificial social intelligence. *Annual review of sociology*, 20(1):407–436.

Chen, H., Sun, M., Tu, C., Lin, Y., and Liu, Z. (2016). Neural sentiment classification with user and product attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1650–1659, Austin, Texas, November. Association for Computational Linguistics.

Frege, G. (1948). Sense and reference. *The philosophical review*, 57(3):209–230.

Heise, D. R. (1987). Affect control theory: Concepts and model. *Journal of Mathematical Sociology*, 13(1-2):1–33.

Heise, D. R. (2010). *Surveying cultures: Discovering shared conceptions and sentiments*. John Wiley & Sons.

Hutto, C. J. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.

Irsoy, O. and Cardie, C. (2014). Opinion mining with deep recurrent neural networks. In *EMNLP*, pages 720–728.

Joseph, K. (2016). *New methods for large-scale analyses of social identities and stereotypes*.

Jovanoski, D., Pachovski, V., and Nakov, P. (2016). On the impact of seed words on sentiment polarity lexicon induction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1557–1567.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Long, Y., Xiang, R., Lu, Q., Huang, C.-R., and Li, M. (2019). Improving attention model based on cognition grounded data for sentiment analysis. *IEEE Transactions on Affective Computing*.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and 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*, pages 142–150. Association for Computational Linguistics.

Melville, P., Gryc, W., and 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*, pages 1275–1284. ACM.

Osgood, C. E., Suci, G. J., and Tannenbaum, P. H. (1964). *The measurement of meaning*. University of Illinois Press.

Pennington, J., Socher, R., and Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. In *Proceedings of EMNLP*, pages 1532–1543.

Qian, Q., Huang, M., Lei, J., and Zhu, X. (2016). Linguistically regularized lstms for sentiment classification. *arXiv preprint arXiv:1611.03949*.

Smith-Lovin, L. and Heise, D. R. (1988). *Analyzing Social Interaction: Advances in Affect Control Theory*, volume 13. Taylor & Francis.

Socher, R., Pennington, J., Huang, E. H., Ng, A. Y., and Manning, C. D. (2011). Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 151–161. Association for Computational Linguistics.

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., and 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)*, volume 1631, page 1642. Citeseer.

Tang, D., Qin, B., and 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*, pages 1422–1432.

Tang, D., Qin, B., and Liu, T. (2015b). Learning semantic representations of users and products for document level sentiment classification. In *Proc. ACL*.

Teng, Z., Vo, D.-T., and Zhang, Y. (2016). Context-sensitive lexicon features for neural sentiment analysis. In *EMNLP*, pages 1629–1638.

Wilson, T., Wiebe, J., and 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*, pages 347–354. Association for Computational Linguistics.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and 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*.

Zou, Y., Gui, T., Zhang, Q., and Huang, X. (2018). A lexicon-based supervised attention model for neural sentiment analysis. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 868–877.