Are Manually Prepared Affective Lexicons Really Useful for Sentiment Analysis

Minglei Li, Qin Lu and **Yunfei Long**
{csmli,csqinlu,csylong}@comp.polyu.edu.hk

Abstract

In this paper, we investigate the effectiveness of different affective lexicons through sentiment analysis of phrases. We examine how phrases can be represented through manually prepared lexicons, extended lexicons using computational methods, or word embedding. Comparative studies clearly show that word embedding using unsupervised distributional method outperforms manually prepared lexicons no matter what affective models are used in the lexicons. Our conclusion is that although different affective lexicons are cognitively backed by theories, they do not show any advantage over the automatically obtained word embedding.

1 Introduction

Sentiment analysis aims to infer the polarity expressed in a text, which has important applications for data analysis, such as product review (Pang et al., 2008), stock market performance (Nguyen and Shirai, 2015), and crowd opinions (Rosenthal et al., 2015). Sentiment lexicons play a critical role in sentiment analysis (Hutto and Gilbert, 2014). A sentiment lexicon contains a list of words with sentiment polarity (positive or negative) or polarity intensity, such as the NRC Hashtag Lexicon (Mohammad et al., 2013) and VADER sentiment lexicon (Hutto and Gilbert, 2014). However, sentiment lexicons may fail for compositional methods to obtain sentiment of larger text units, such as phrases and sentences. For example, the phrase *avoid imprisonment* expresses positive sentiment. However, when we use sentiment lexicon, it is hard to classify this phrase because both *avoid* and *imprisonment* are negative in both VADER (Hutto and Gilbert, 2014) and NRC Hasntag (Mohammad et al., 2013) lexicons.

In addition to polarity based sentiment lexicons, which can be considered as one-dimensional affective lexicons, different multi-dimensional affect models are also proposed to represent affective information of words, such as the evaluation-potency-activity (EPA) model (Osgood, 1952) and the valence-arousal-dominance (VAD) model (Ressel, 1980). Sentiment can be seen as one of the dimensions under these affective models, such as the evaluation dimension of EPA, and the valence dimension of VAD. Aside from the EPA based lexicon (Heise, 2010), VAD based lexicons include ANEW (Bradley and Lang, 1999), extended ANEW (Warriner et al., 2013), and CVAW (Yu et al., 2016). Although multi-dimensional affective lexicons are theoretically sound, there are mainly two issues. The first one is how to obtain good coverage for affective lexicons. The second one is how to infer the representation of larger text units using word information in the affective lexicons. A previous work uses the average value of the component words as the final representation of larger texts (Yu et al., 2016).

Word embedding has recently been used to represent word semantics, such as word2vec (Mikolov et al., 2013) and Glove (Pennington et al., 2014). Word embedding represents a word as a dense vector, which can be used to measure semantic similarity of words more accurately.

To infer the representation of larger text units based on word embedding, different composition models are proposed, such as weighted addition and multiplication (Mitchell and Lapata, 2008), tensor product (Zhao et al., 2015), recursive neural network (Socher et al., 2013), recurrent neural network (Irsoy and Cardie, 2014), and convelutional neural network (Kim, 2014). Attempts have also been made to infer the affective labels

Proceedings of the The 8th International Joint Conference on Natural Language Processing, pages 146–150,
Taipei, Taiwan, November 27 – December 1, 2017 ©2017 AFNLP

of phrases based on the VAD model using compositional methods (Palogiannidi et al., 2016). However, between the VAD representation and word embedding, it is not clear which one is more effective for sentiment analysis.

Sentiment lexicons, multi-dimensional affective lexicons, and word embedding all represent a word with semantic information. Other than word embedding, all the other lexicons are specifically built for sentiment/affective analysis. Although these representations can be used for sentiment analysis of larger text units, there is no systematic comparison to test their effectiveness. In this paper, we investigate whether the manually annotated sentiment/affective lexicons have some advantage over automatically obtained word embedding on sentiment analysis tasks. Our approach is to use different word level representations to predict the sentiment of phrases to determine which representation of words is more effective. Experiments clearly show that word embedding outperforms manual affective lexicons and extended affective lexicons.

2 Related Work

To apply a sentiment lexicon in sentiment analysis, the simpliest way is to take word present in a lexicon as a simple feature (Pang et al., 2008). For intensity-based sentiment lexicons, the sentiment value can be aggregated by addition of every sentiment linked word in a sentence (Hutto and Gilbert, 2014; Vo and Zhang, 2016). Another method is to use sentiment related features, such as total count of sentiment tokens, total sentiment score, maximal sentiment score, etc.(Mohammad et al., 2013; Tang et al., 2014).

Many efforts have been made to construct multi-dimensional affective lexicons, such as ANEW for English (Bradley and Lang, 1999; Warriner et al., 2013), CVAW for Chinese (Yu et al., 2016), and other languages (Montefinese et al., 2014; Imbir, 2015). However, only few works use multi-dimensional affective lexicons for affective analysis. The work by Yu et al. (2016) uses the average VAD values of individual words as the VAD value of a sentence. In (Palogiannidi et al., 2016), affective representation of phrases is obtained through matrix-vector multiplication, where modifier words are represented by matrices and head words are represented as VAD vectors.

When word embedding is used for sentiment analysis, different composition methods are used to infer the representation of a sentence, such as simple addition, weighted addition (Mitchell and Lapata, 2008), recurrent neural networks (Irsoy and Cardie, 2014), and convolutional neural networks (Kim, 2014).

However, there is no systematic comparison between lexicon based representations and word embedding representations for sentiment analysis. This is the motivation of our work.

3 Comparison Method

Our objective is to study the effectiveness of different word representations for units longer than words for sentiment analysis. To focus more on the effectiveness of representations, we only study bigram phrases in this paper. The following is the list of lexicon resources and embeddings used for this comparative study.

1. The VADER sentiment lexicon of size 7,502, annotated through crowdsourcing (Hutto and Gilbert, 2014). Its value range is [-4, 4].

2. The NRC Hasntag sentiment lexicon (denoted as HSenti) of size 54,129, constructed automatically based on hashtags (Mohammad et al., 2013).

3. The multi-dimensional EPA lexicon of size 2,000, annotated manually (Heise, 2010) in three dimensions of evaluation, potency, activity in the range of [-4.3, 4.3].

4. The multi-dimensional VAD lexicon of size 13,915, annotated through crowdsourcing (Warriner et al., 2013). The annotation is in three dimensions of valence, arousal, dominance in the range of [1, 9].

5. Word embedding of 300 dimension with size of 2,196,017 trained by the the Glove model using unsupervised matrix factorization on a corpus of size 840 billion (Pennington et al., 2014), denoted as g300.

The manually annotated lexicons have limited sizes. For fair comparison, we use the state-of-the-art method proposed by Li et al. (2016), which train a Ridge regression model using word embedding as features to automatically extend the manually constructed lexicons so that all the vocabularies of different lexicons are the same size of g300.

Let us use the term base representations to refer to the different word representations used in this comparative study. We first construct the representation of a phrase from the base representations of its component words using some composition functions. Then, we perform sentiment prediction for phrases to evaluate which of the base representations is more effective.

In a composition model, the representation of a phrase is inferred from that of its component words. Given a phrase p with two component words w^1 and w^2 and their respective base representations $\vec{w}^1$ and $\vec{w}^2$, the representation of p, denoted by $\vec{p}$, can be constructed by a function f:

$$\vec{p} = f(\vec{w}^1, \vec{w}^2). \tag{1}$$

Different composition models are proposed for f (Mitchell and Lapata, 2008). An addition composition model can be defined as

$$\vec{p} = \vec{w}^1 + \vec{w}^2. \tag{2}$$

The multiplication composition model is defined by element-wise vector multiplication:

$$\vec{p} = \vec{w}^1 * \vec{w}^2. \tag{3}$$

The concatenation composition function that simply concatenates the two vectors:

$$\vec{p} = [\vec{w}^2, \vec{w}^1]. \tag{4}$$

A more advanced composition model is the Recurrent Neural Network (RNN). Here we use the most widely used Long Short Term Memory (LSTM) network (Hochreiter and Schmidhuber, 1997) as our composition model. It models a word sequence as:

$$\vec{i}_t = \sigma(U_i\vec{x}_t + W_i\vec{h}_{t-1} + \vec{b}_i), \tag{5}$$

$$\vec{f}_t = \sigma(U_f\vec{x}_t + W_f\vec{h}_{t-1} + \vec{b}_f), \tag{6}$$

$$\vec{o}_t = \sigma(U_o\vec{x}_t + W_o\vec{h}_{t-1} + \vec{b}_o), \tag{7}$$

$$\vec{q}_t = tanh(U_q\vec{x}_t + W_q\vec{h}_{t-1} + \vec{b}_q), \tag{8}$$

$$\vec{p}_t = \vec{f}_t * \vec{p}_{t-1} + \vec{i}_t * \vec{q}_t, \tag{9}$$

$$\vec{h}_t = \vec{o}_t * tanh(\vec{p}_t). \tag{10}$$

Here $\vec{x}_t$ is the representation of an input word representation at step t, $\vec{w}^1$ or $\vec{w}^2$. $\vec{i}_t$, $\vec{f}_t$, $\vec{h}_t$, $\vec{p}_t$, $\vec{q}_t$ are internal representations and $\vec{o}_t$ is current output representation. U_i, U_f, U_o, U_q are the model matrix parameters. Sentiment prediction is performed on the output representation the final step.

In this work, we use different composition functions to evaluate the effectiveness of different base representations.

4 Evaluation on the Comparisons

The five lexicons introduced in Section 3 are used for evaluations.

4.1 Experiment Setting

For comparison, we first extracted a set of phrases with sentiment ratings from the Stanford Sentiment Treebank (SST) (Socher et al., 2013), in which every sentence is parsed and each node in the parsed tree has a sentiment score ranging between [0, 1], and obtained by crowdsourcing. We only extract adjective-noun phrases, noun-noun phrases and verb-noun phrases, and the size of the final phrase collection in SST is 9,922. Note that only 6,736 words are used for this set of phrases and they are present in all the five lexicons used.

Based on this phrase set, we construct three sentiment analysis tasks: (1) a regression task to predict the sentiment score of phrases (labeled as **SST-R**); (2) a binary classification task by converting sentiment scores to discrete labels, where positive label is no less than 0.6 and negative label is no more than 0.4 (labeled as **SST-2c**); (3) a ternary classification task similar to SST-2C except that there is an addition of neutral label in the range of 0.4-0.6 (labeled as **SST-3c**).

Different evaluation metrics are used for the three different tasks. Mean absolute error (mae) and Kendall rank correlation coefficient (τ) are used for SST-R. Accuracy and F-score are used for SST-2c. Weighted accuracy and weighted F-score are used for SST-3c. Ridge regression and SVM with the linear kernel are used for regression and classification task, respectively[1]. For LSTM, the output layer is set differently for regression and classification tasks respectively [2]. The number of hidden dimensions in LSTM is set to 4. In all the experiments, 5-fold cross validation is used. Results are based on the best parameters we can obtain in our experiments.

4.2 Result and Analysis

Table 1 shows the result of the three tasks. Let us first take a look at the different composition functions. Multiplication performs the worst in all categories. On the other hand, LSTM, as a deep learning method, is the best performer. Addition and concatenation do have comparable performance and not too off from LSTM

[1] Using the scikit-learn tool: scikit-learn.org/

[2] Using the Keras tool: https://keras.io/

Feature	Comp	SST-R		SST-2c		SST-3c	
		mae	τ	**acc**	**f**	**acc**	**f**
VADER	mul	0.103	0.240	0.664	0.786	0.608	0.507
VADER	add	0.088	0.477	0.889	0.913	0.643	0.578
VADER	conc	0.086	0.482	0.888	0.912	0.655	0.591
VADER	lstm	0.086	0.487	0.894	0.917	0.667	0.655
HSenti	mul	0.110	0.060	0.636	0.777	0.573	0.418
HSenti	add	0.102	0.298	0.766	0.826	0.573	0.418
HSenti	conc	0.102	0.304	0.768	0.829	0.573	0.418
HSenti	lstm	0.100	0.307	0.770	0.825	0.610	0.556
EPA	mul	0.097	0.367	0.833	0.871	0.575	0.420
EPA	add	0.092	0.422	0.888	0.913	0.600	0.488
EPA	conc	0.092	0.427	0.887	0.912	0.602	0.493
EPA	lstm	0.091	0.438	0.893	0.915	0.633	0.605
VAD	mul	0.089	0.456	0.897	0.919	0.618	0.544
VAD	add	0.090	0.451	0.890	0.913	0.620	0.549
VAD	conc	0.089	0.459	0.894	0.917	0.625	0.557
VAD	lstm	0.090	0.466	0.891	0.915	0.635	0.602
g300	mul	0.106	0.246	0.635	0.777	0.575	0.420
g300	add	0.074	0.564	0.923	0.939	**0.755**	**0.749**
g300	conc	0.073	0.565	0.920	0.937	0.754	0.748
g300	lstm	**0.070**	**0.573**	**0.926**	**0.941**	0.751	**0.749**

Table 1: Performance of different word representations under different composition functions for SST phrase sentiment analysis. mul: multiplication composition. add: addition composition. conc: concatenation composition.

on SST-R and SST-2c. Secondly, for the two sentiment lexicons, VADER performs much better than HSenti lexicon. This may be because that VADER is manually annotated from crowdsourcing whereas HSenti is automatically obtained which contains more noise. Thirdly, for the two multi-dimensional affective lexicons, VAD performs slightly better than EPA. It is surprising that the multi-dimensional lexicons perform even worse than the sentiment lexicon VADER even though the annotated size of VAD (13,915) is much larger than VADER (7,502). This puts a question mark on the quality of annotation for multi-dimensional lexicon resources. Fourthly, word embedding[3] performs much better than all the other representations. For instance, it achieves a relative improvement of 17.7% under τ for SST-R over the secondly ranked VADER representation. Different composition functions for word embedding perform comparably. In principle, LSTM would have more benefits if the text length is longer. In this study, the performance difference is not obvious because our phrases are only bigrams.

In the first experiment, manually constructed affective lexicons are extended for comparison to be performed on the same set of word list. Since automatically extended lexicons can introduce errors, we perform the second experiment using only a manually annotated lexicon. We use the largest original VAD lexicon without extension to compare with word embedding. In this case, the intersection of VAD and word embedding has 3,908 words. The subset corpus of SST containing these words has 5,251 phrases. We perform 5-fold cross validation on this dataset. The result is shown in **Table 2**. Again, word embedding achieves much better result than manually annotated VAD lexicon. If coverage issue is considered, word embedding has even more advantages. Interestingly, comparison between **Table 1** and **Table 2** shows that the manually annotated lexicon does not perform better than its automatically extended lexicon even without considering the coverage problem.

Feature	Comp	SST-R$_s$		SST-2c$_s$		SST-3c$_s$	
		mae	τ	**acc**	**f**	**acc**	**f**
VAD	add	0.093	0.450	0.901	0.927	0.614	0.554
VAD	conc	0.092	0.465	0.905	0.931	0.624	0.568
VAD	lstm	0.093	0.471	0.885	0.916	0.620	0.594
g300	add	0.075	0.575	0.926	0.945	0.759	0.754
g300	conc	0.075	0.575	0.931	0.949	**0.762**	**0.757**
g300	lstm	**0.071**	**0.588**	**0.934**	**0.951**	0.754	0.753

Table 2: Performance of manually annotated VAD and corresponding word embedding representations under different composition functions for phrase sentiment analysis.

5 Conclusion

Automatically obtained word embedding clearly outperforms both manually and automatically obtained affective lexicons including sentiment lexicons and multi-dimensional affective lexicons. Although different affective models are backed by cognitive theories and affective lexicons are designed specifically for affective analysis, building them consumes too much resources and annotation quality may still be questioned due to added complexity. Through a downstream task of sentiment labeling of phrases, we conclude that the manually annotated affective lexicons have no advantage over word embedding under different composition models. However, affective lexicons as resources can still be used as additional features rather than being used alone.

[3] We also experiment on different word embedding dimensions including 50,100,200. All are better than the other lexicons.

Acknowledgments

This work is supported by HK Polytechnic University (PolyU RTVU and CERG PolyU 15211/14E).

References

Margaret M Bradley and Peter J Lang. 1999. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Citeseer.

David R Heise. 2010. *Surveying cultures: Discovering shared conceptions and sentiments*. John Wiley & Sons.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International AAAI Conference on Weblogs and Social Media*.

Kamil K Imbir. 2015. Affective norms for 1,586 polish words (anpw): Duality-of-mind approach. *Behavior research methods*, 47(3):860–870.

Ozan Irsoy and Claire Cardie. 2014. Modeling compositionality with multiplicative recurrent neural networks. *arXiv preprint arXiv:1412.6577*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Minglei Li, Yunfei Long, and Qin Lu. 2016. A regression approach to valence-arousal ratings of words from word embedding. In *International Conference on Asian Language Processing*, pages 120–123. IEEE.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *ACL*, pages 236–244.

Saif M Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242*.

Maria Montefinese, Ettore Ambrosini, Beth Fairfield, and Nicola Mammarella. 2014. The adaptation of the affective norms for english words (anew) for italian. *Behavior research methods*, 46(3):887–903.

Thien Hai Nguyen and Kiyoaki Shirai. 2015. Topic modeling based sentiment analysis on social media for stock market prediction. In *ACL*, pages 1354–1364.

Charles E Osgood. 1952. The nature and measurement of meaning. *Psychological bulletin*, 49(3):197.

Elisavet Palogiannidi, Elias Iosif, Polychronis Koutsakis, and Alexandros Potamianos. 2016. A semantic-affective compositional approach for the affective labelling of adjective-noun and noun-noun pairs. In *WASSA@NAACL-HLT*.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.

JA Ressel. 1980. A circumplex model of affect. *J. Personality and Social Psychology*, 39:1161–78.

Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif M Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. Semeval-2015 task 10: Sentiment analysis in twitter. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 451–463.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, Christopher Potts, et al. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, volume 1631, page 1642. Citeseer.

Duyu Tang, Furu Wei, Bing Qin, Ming Zhou, and Ting Liu. 2014. Building large-scale twitter-specific sentiment lexicon: A representation learning approach. In *COLING*, pages 172–182.

Duy Tin Vo and Yue Zhang. 2016. Dont count, predict! an automatic approach to learning sentiment lexicons for short text. In *Proceedings of ACL*, volume 2, pages 219–224.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207.

Liang-Chih Yu, Lung-Hao Lee, Shuai Hao, Jin Wang, Yunchao He, Jun Hu, K Robert Lai, and Xuejie Zhang. 2016. Building chinese affective resources in valence-arousal dimensions. In *Proceedings of NAACL-HLT*, pages 540–545.

Yu Zhao, Zhiyuan Liu, and Maosong Sun. 2015. Phrase type sensitive tensor indexing model for semantic composition. In *AAAI*, pages 2195–2202.