Retrofitting of Pre-trained Emotion Words with VAD-dimensions and the Plutchik Emotions

Manasi Kulkarni[1,2]
[2]Computer Engineering and IT
Veermata Jijabai Technological Institute
Mumbai
manasi@cse.iitb.ac.in

Pushpak Bhattacharyya[1]
[1]Computer Science and Engineering
Indian Institute of Technology Bombay
Mumbai
pb@cse.iitb.ac.in

Abstract

The word representations are based on distributional hypothesis according to which words that occur in the similar contexts, tend to have a similar meaning and appear closer in the vector space. For example, the emotionally dissimilar words "joy" and "sadness" have higher cosine similarity. The existing pre-trained embedding models lack in emotional words interpretations. For creating our VAD-Emotion embeddings, we modify the pre-trained word embeddings with emotion information. This is a lexicons based approach that uses the Valence, Arousal and Dominance (VAD) values, and the Plutchik's emotions to incorporate the emotion information in pre-trained word embeddings using post-training processing. This brings emotionally similar words nearer and emotionally dissimilar words away from each other in the proposed vector space. We demonstrate the performance of proposed embedding using NLP downstream task - Emotion Recognition.

1 Introduction

An emotion is a feeling that characterizes the state of mind such as happiness, sadness, anger, fear and more. The emotions are classified using various taxonomies under the dimensional models and the psychological emotion models such as Ekman (1992) Emotion Model , Plutchik (1980) Emotion Wheel, Parrot (2001) Model which agree to a basic set of emotion with few changes. The Plutchik emotion model represents the categorization of emotion words into 8 basic emotions : anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. The PAD emotional state model (Mehrabian, 1994) is a 3-dimensional model that represents every emotion in Valence (Pleasure), Arousal and Dominance dimensions.

Emotion detection in the text is critical for a number of applications and services in diverse domains, including market research, customer-care, psychological healthcare, and intelligent tutoring systems and so on.(Mohammad and Turney, 2013). The automatic detection of emotions remains a challenging task till date as researchers may use different emotion models with different number and types of emotion categories. Also, the emotions are subjective, hence creation of emotion related resources requires much time and effort.

Word embedding are distributed word representations where each word w in the vocabulary V is mapped into a dense, low-dimensional, continuous valued vector $v_w \epsilon R^d$. Here d represents dimensions of the vector space model. Most of the embeddings are modeled using the syntactic context of words which means words appearing in the similar contexts have the similar semantics and appear closer in the vector space (Mikolov et al., 2013), (Pennington et al., 2014). As a consequence, emotionally opposite words, such as "joy" and "sorrow" occurring in similar contexts show higher cosine similarity. Hence, said property does not fit in case of emotion words as 'joy' and 'sorrow' and many more similar and opposite emotion words too.

We propose a model that modifies the pre-trained word embedding with emotion information using a post-training processing method. This is a lexicons based approach that uses the Valence, Arousal and Dominance (VAD) values, and the Plutchik's emotions to incorporate the emotion information in the word embeddings.

The main contributions of this paper includes the creation of emotion-fitted embedding to be used for NLP downstream tasks related to emotion analysis. We present average of cosine similarities for emotion words. The visualizations shown for NRC-emolex lexicons using for Glove-300d

embeddings, and retrofitted embeddings at both steps : VAD-append embeddings and VAD-Emotion embeddings confirms the step-by-step clustering of similar emotion words. The accuracy for emotion recognition as downstream emotion task using the proposed embedding is mentioned as results. Though clustering of emotion words takes place, the accuracy for emotion recognition task using proposed embedding is closer to baseline but can not outperform the same.

The paper is organized as follows. Section 2 discusses various approaches used by researchers for updating vector space models for specific set of NLP tasks. Section 3 describes the proposed approach. experiment and setting are explained in section 4. We discuss the results and observation in section 5.

2 Related Work

The use of lexical semantic information (lexical resources), sentiment information, emotion information to improve distributional representations in respective area of NLP tasks is recent. Methods like Tang et al. (2016), Agrawal et al. (2018), Ye et al. (2018) achieve improved representations by using training on unlabelled corpora, distant supervision and other techniques to gain relational knowledge to modify the prior or add a regularization term. Such methods are known as 'pre-training methods', as they alter the training process for word representations. Such methods may require a change in the loss function with training and may be computationally expensive.

Ye et al. (2018) proposed a method for sentiment analysis, where they use external knowledge from SentiWordNet (Baccianella et al., 2010), Extended ANEW lexicons (Warriner et al., 2013) with pre-trained word embeddings during joint parameter training to a CNN classifier using training data. Agrawal et al. (2018) proposed a distant supervision method for automatically labeling a large corpus of training data with fine-grained emotions; and the LSTM model architectures for learning emotion-enriched word embedding from this training data.

On the other hand post-training methods include external information to modify to the vanilla word representations such as Word2Vec, GLOVE to name a few. Retrofitting Method (Faruqui et al., 2015), has used word relation knowledge from semantic lexicons (e.g. WordNet), to bring similar words closer in the retrofitted vector space. It injects antonym and synonym constraints to improve the existing word representations. Mrkšić et al. (2016) presented post-training approach named as counter-fitting which injects antonym and synonym constraints into existing vector space representations in order to improve the vectors' capability for increase semantic similarity.

Aff2Vec (Khosla et al., 2018) aims at incorporating affective information in word representations. They have used the Warriner's VAD lexicon (Warriner et al., 2013) to improve the strength in the antonym-synonym relationships of the words is incorporated to the word distribution space. The word similarity task and other sentiment analysis related tasks are performed to display the results.

Seyeditabari et al. (2019) proposed a method, based on counterfitting approach (Mrkšić et al., 2016) that incorporates emotional information of words into the model. It uses an NRC-emotional lexicon (Mohammad and Turney, 2013) and the Plutchik's model of basic emotions to prepare emotion constraints for fitting an emotional information into pre-trained word vectors.

Here, we propose a pipeline model to integrate Valence, Arousal and Dominance information of emotion words and their basic emotion labels from lexicon set to create a new vector space on pre-trained word vectors using post-training (post-processing) method.

3 Fitting VAD values and Emotion constraints into the Word Embedding

This work aims at incorporating affect information in word representations. The figure - 1 illustrates overview of proposed system for retrofitting of pre-trained vector space. The emotion information is infused in two consecutive steps. The subsections 3.1 and 3.2 discuss the processing performed in step-1 and step-2 respectively. As shown in figure-1, at step-1, we append the Valence, Arousal and Dominance (VAD) values for the emotion lexicons to their respective word embedding to transform an original pre-trained vector space V into the VAD-appended vector space V'. In step-2, we use emotion constraints to modify these VAD-appended embeddings further to achieve retrofitted

vector space V". Following subsections present the steps of modifications in the pre-trained word vector space V to V' and then to the final modified vector space V" from V'.
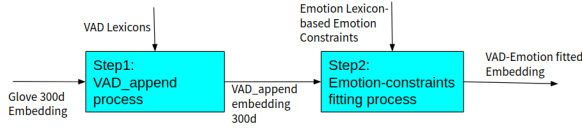

Figure 1: The System Overview

NRC-VAD lexicons : NRC-VAD lexicon (Mohammad, 2018) is a set of affect lexicons with 20007 English words. It contains real-valued scores for valence, arousal, and dominance (VAD) on a scale of [0 - 1] for each lexicon.

3.1 Step-1 : VAD-Append

Consider the word embedding space V and the affect ([V,A,D]) embedding space A. The word vector v_w of word w , $v_w \epsilon R^M$, is concatenated with the respective VAD-vector $a_w = [V, A, D] \epsilon R^3$ from A, resulting in a $M + 3$ dimensional word representation (Khosla et al., 2018). For the words not present in VAD-NRC lexicons, $a_w = [0.5, 0.5, 0.5] \epsilon R^3$ is assumed. Then these vectors are reduced to M dimensions using dimensionality reduction algorithm such as Principle Component Analysis (PCA) so that their performance can be compared with existing pre-trained embedding. This process transforms a pre-trained vector space V to VAD-appended vector space V' as shown in figure-2.

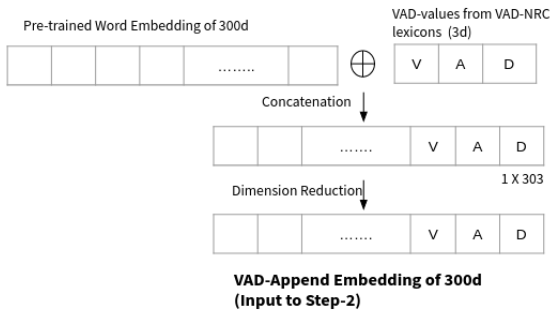**VAD-Append Embedding of 300d (Input to Step-2)**

Figure 2: The VAD-Append Process for a word

3.2 Step-2 : Emotion constraints using the Plutchik Model of emotions

For fitting emotional information into VAD-Append word vectors, we use a methodology on

Emotion-1	Emotion-2	Differing dimension from VAD
Anger	Fear	D
Anticipation	Surprise	AD
Disgust	Trust	VD
Joy	Sadness	VAD

Table 1: VAD-dimensions and opposite emotions

the similar lines of (Seyeditabari et al., 2019). We aim to modify VAD-Append vector space $V' = \{v_1', v_2', \ldots, v_n'\}$ to new vector space $V" = \{v_1", v_2", \ldots, v_n"\}$ to add emotion information without loosing much information present with original vectors.

The Plutchik's emotion model defines 4 pairs of opposite emotions: Anger and Fear, Disgust and Trust, Anticipation and Surprise, and Joy and Sadness. These emotions differ based on high(1) or low(0) VAD-values for the respective emotions. Anger and Fear differs on Dominance value as Dominance is high for Anger and low for Fear. Table-1 shows difference for rest of the emotions.

We refine the VAD-append vectors further to increase the cosine similarity between words with similar emotions and decrease cosine similarity between the dissimilar emotion words which can help improve new vector space having word vectors with interpretation of emotions in the respective words.

To achieve this, two emotion constraint lists are created. First list $TrueEmotion$ maintains pairs as (word, true_emotion) for every lexicon from NRC-Emolex such as $\{(w_1, e_1), (w_1, e_2) \ldots (w_2, e_1), \ldots (w_n, e_3) \ldots\}$ and another list $OppositeEmotion$ maintains pairs as (word, opposite_emotion) for every pair present in the first list such as $\{(w_1, o_1), (w_1, o_2) \ldots (w_2, o_1) \ldots (w_n, o_3) \ldots \}$. Here o_i represents the opposite emotion of ith emotion e_i as shown on the Plutchik wheel of emotions. The NRC-Emolex lexicons (Mohammad and Turney, 2013) are annotated with the best suitable Plutchik emotions and positive or negative as sentiment value are used for the same.

Our objective function for step-2 is based on the counterfitting approach (Mrkšić et al 2016) to decrease the cosine distance between words with their associated emotion in the list

$TrueEmotion(TE)$, and to increase cosine distance with their opposite emotions in the list $OppositeEmotion(OE)$. The objective function is to be minimised to achieve proposed vector space model V'. The objective function contains following three terms,

$$Obj(V', V'') = c_1 OR(V'') + c_2 TA(V'') \\ + c_3 VSP(V', V'') \qquad (1)$$

Opposite Repels(OR) : This term is to reduce cosine similarity between words' and opposite emotions' vectors away from each other in by increasing cosine distance between them in the transformed vector space V'. It uses the $OppositeEmotion$ list for this purpose.

$$OR(V'') = \sum_{(w,o)\epsilon OE} max(0, \delta - d(v_w'', v_o'')) \qquad (2)$$

Here, standard value of $\delta = 1$ and cosine distance $d(v_w, v_o) = 1 - cos_{dist}(v_w, v_o)$

True Attracts(TA): This term is to bring the embeddings of the words and their respective true emotions nearer to each other in new vector space V'. In other words, for increasing cosine similarity between them.

$$TA(V'') = \sum_{(w,e)\epsilon TE} max(0, d(v_w'', v_e'') - \gamma)) \qquad (3)$$

The $\gamma = 0$ represents minimum distance between true emotion and words.

Vector Space Preservation(VSP) : The original vector space describes the distributional information for words from very large textual corpora (Mrkšić et al., 2016). VSP term tries to minimize the difference between cosine distance between word pairs in original vector space V and new vector space V' to preserve semantic and contextual information as much as possible. In current experiments, the neighbouring words are chosen from NRC-Emolex only.

$$VSP(V', V'') = \\ \sum_{i=1}^{N} \sum_{j\epsilon N(i)} (max(0, d(v_x'', v_y'') - d(v_x', v_y'))) \qquad (4)$$

.

4 Experiments

We have used the NRC-VAD lexicons (Mohammad, 2018) to create VAD vectors and the NRC-Emolex emotion lexicons (Mohammad and Turney, 2013) for emotion constraints creation as they are labelled with Plutchik's wheel of emotions (Plutchik, 1980) along with positive and negative sentiment values.

During experiments, Glove-300d (Pennington et al., 2014), FastText-1M (Joulin et al., 2016), and Google Word2Vec (Mikolov et al., 2013) are used as pre-trained input vector space and created the VAD-append embeddings at step-1, VAD-Emotion word embeddings at step-2 respectively. We run Stochastic Gradient Descent (SGD) for 20 epochs to achieve the final retrofitted embeddings with emotion information.

We have compared performance of VAD-Emotion fitted embeddings with existing vector space as mentioned above for the task of finding average cosine similarity over NRC-Emotion lexicons for similar and opposite emotions. Also, have displayed accuracy results of Emotion Recognition task on ISEAR dataset (ISEAR) and the Twitter Emotion Corpus (TEC) dataset (Mohammad, 2012).

5 Results and Discussion

Table-2 shows average cosine similarity between NRC-Emolex lexicons and their respective emotions. Higher the cosine similarity is better for the similar emotion words. For Joy and Trust emotions, VAD-Emotion Embedding perform better than rest of them.

Table-3 shows average cosine similarity between NRC-Emolex lexicons and the opposite emotions.The cosine similarity between lexicons and emotion should be as low as possible. The range for cosine similarity is $[+1, -1]$ The figure - 5 shows the clustering of similar emotion words together and opposite emotion words in opposite word clusters.

Table-4 shows accuracy of Emotion Recognition Task using the pre-trained word embedding, and respective VAD-append and VAD-Emotion Embedding as input to a simple BiLSTM model. The Table-4 reports 4-fold cross validation accuracy on the subset of ISEAR dataset with

Figure 3: Visualization of NRC-Emolex lexicons using Glove-300d embeddings

Figure 4: Visualization of NRC-Emolex lexicons using VAD-Append embeddings after step-1 processing

Figure 5: Clustering of NRC-Emolex lexicons using VAD-dimention + Plutchik emotion embeddings after step-2 processing

5.4k examples and TEC dataset with 20k examples respectively used for training and testing.

For the purpose of comparison, the emotion recognition task on the ISEAR dataset was performed with the Emotion-Refined-Embedding (Seyeditabari et al., 2019) with Glove-300d pre-trained embedding which resulted as accuracy of **64.84%**. The embeddings are computed with help of the code of Seyeditabari et al. (2019) which is available publicly.

It can be observed from emotion recognition accuracy values that VAD-Append embedding give better accuracy than pre-trained Embedding - Glove, Google Word2Vec, FastText as well as respective VAD-Emotion embedding. The initial segregation based on V,A,D values helps to achieve improvement in accuracy of emotion recognition.

The figure-3, figure-4, and figure-5 show visualization for NRC-Emolex lexicons using Glove-300d embeddings, VAD-append embeddings (step-1 output) and VAD-Emotion embeddings i.e. final proposed retrofitted embeddings. It can be observed that VAD-Emotion embeddings show the better cosine similarity among the emotionally similar words and less cosine similarity between emotionally dissimilar words than rest of them. This can be confirmed by looking at the clusters of emotion lexicons formed in figure - 5. Yet, the accuracy for emotion recognition, do not show better results for VAD-Emotion Embedding.

The reason from the primary observation is that it is due to overlaps in clusters, as one lexicon may belong to one or more emotions. Also, at step-2, we

retrofit word embeddings, only for emotion words present in NRC-Emolex. The limitation with post-processing methods such as counter-fitting is that it retrofits emotion words present in the constraints. Hence, we should perform a global specialization or post-specialization processing (Vulic et al., 2018) for retrofitting non-emotion words with reference to the newly retrofitted emotion word vectors. This will turn into the retrofitting of complete vector space with the Plutchik's emotion information, which may help to improve accuracy of the downstream tasks.

At step-1 every word embedding is appended with VAD-values or the neutral value vector [0.5,0.5,0.5] which retrofits every word for VAD-values in V' vector space. Due to the limitation mentioned above, step-2 does not perform better than step-1. Hence, the words which are not in the NRC-lexicons can not be retrofitted at step-2. To achieve the better accuracy results with final retrofitted vector space V", this is going to be our future work with the proposed method.

Conclusion and Future Work

Embedding models have an important role in word representation in various natural language processing tasks. Here we present an approach to bring emotionally similar words nearer and emotionally dissimilar words away from each other in the proposed vector space. This can be observed in the better cosine similarity results presented in table - 2, 3 for the NRC-Emolex lexicons and also in the in figure - 5 that shows the similar emotion words from Emolex are clustered together and opposite emotion words are far apart. In this

Lexicons labelled with Emotion	GloVe-300d Embedding	VAD-append Glove Embedding (Step-1 of Proposed Approach)	VAD-Emotion Embedding (Step-2 of Proposed Approach)
Anger	0.2864	0.4650	0.2548
Anticipation	0.3746	0.5496	**0.5743**
Disgust	0.2789	0.4436	**0.7646**
Fear	0.3057	0.4945	0.4023
Joy	0.3315	0.4585	**0.7720**
Sadness	0.2603	0.4329	**0.7214**
Surprise	0.2818	0.4287	0.3906
Trust	0.2567	0.4246	**0.8724**

Table 2: Average on Cosine Similarity between lexicons and their emotion labels from NRC-Emolex

Lexicons labelled with Emotion	Opposite Emotion	GloVe-300d Embedding	VAD-append Glove Embedding (Step-1 of Proposed Approach	VAD-Emotion Embedding (Step-2 of Proposed Approach)
Anger	Fear	0.3091	0.4329	**0.0181**
Anticipation	Surprise	0.2519	0.3528	**-0.0235**
Disgust	Trust	0.1625	0.1843	**-0.2284**
Fear	Anger	0.2557	0.5078	**-0.0127**
Joy	Sadness	0.2372	0.3118	**-0.1035**
Sadness	Joy	0.2029	0.2719	**-0.1164**
Surprise	Anticipation	0.2422	0.4145	**-0.0223**
Trust	Disgust	0.1626	0.2656	**-0.2264**

Table 3: Average on Cosine Similarity between lexicons from NRC-Emolex and their opposite emotions

Word Embedding	ISEAR dataset			Twitter Emotion Corpus (TEC)		
	Pre-Trained Embedding	VAD-Append Embedding (step-1)	VAD-Emotion Embedding (step-2)	Pre-Trained Embedding	VAD-Append Embedding (step-1)	VAD-Emotion Embedding (step-2)
Glove-300d	70.57	*70.95*	66.11	56.44	*57.62*	54.73
Google Word2Vec	69.7	*70.50*	65.84	56.90	*58.51*	53.81
FastText-1M	67.79	*70.60*	66.11	55.35	*57.70*	55.11

Table 4: Average accuracy of Emotion Recognition model for 4-fold cross validation on ISEAR dataset and TEC dataset

approach, we have combined effect of dimensional emotion model through VAD values as well as effect of psychological emotion model through emotion constraints for incorporating emotion information in the proposed vector space VAD-Emotion embedding.

However, the VAD-append embedding shows better results than pre-trained embedding, the accuracy of VAD-Emotion embedding for emotion recognition task is not impressive. The post-specialization or global-specialization process for retrofitting of non-emotion word embeddings may improve the accuracy at step-2 of proposed model.

As a future work, we will be performing a process for retrofitting non-emotion words with reference to the newly retrofitted word vectors, which may help to improve accuracy of the downstream emotion-based NLP tasks. Also, use of The knowledge base such as ConceptNet, SenticNet etc may help in finding more emotion words and emotion constraints to gain better emotion information. Being the lexicon-based approach, the approach with modifications may be useful for emotion/sentiment based applications with low-resource languages too.

References

Ameeta Agrawal, Aijun An, and Manos Papagelis. 2018. Learning emotion-enriched word representations. In *Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, USA*, pages 950–961.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)*.

Paul Ekman. 1992. An argument for basic emotions. *Cognition Emotion, 6(3-4)*, pages 169–200.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615.

ISEAR. The international survey on emotion antecedents and reactions as a python dataset for machinelearning http://www.affective-sciences.org/researchmaterial.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Paper*, pages 427–431.

Sopan Khosla, Niyati Chhaya, and Kushal Chawla. 2018. Aff2vec: Affect–enriched distributional word representations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2204–2218.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*.

Saif Mohammad. 2012. emotional tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*Sem), Montreal, Canada*.

Saif M Mohammad. 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*, pages 174–184.

Saif M Mohammad and Peter D Turney. 2013. Crowd-sourcing a word–emotion association lexicon. In *Computational Intelligence, 29(3)*, pages 436–465.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. In *Proceedings of NAACL-HLT 2016*, pages 142–148.

W Parrot. 2001. Emotions in social psychology. key readings in social psychology. In *Philadelphia: Psychology Press. ISBN 978-0863776830*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 1532–1543.

Robert Plutchik. 1980. Emotion: Theory, research, and experience. In *Vol. 1. Theories of emotion, 1, New York: Academic*.

Armin Seyeditabari, Narges Tabari, Shefie Gholizade, and Wlodek Zadrozny. 2019. Emotional embeddings: Refining word embeddings to capture emotional content of words. In *arXiv:1906.0011v2, June 2019*.

D. Tang, F. Wei, B. Qin, N. Yang, T. Liu, and M. Zhou. 2016. Sentiment embeddings with applications to sentiment analysis. In *IEEE Transactions on Knowledge and Data Engineering 28, 2(2016)*, pages 496–509.

Ivan Vulic, Goran Glavaš, Nikola Mrkšic, and Anna Korhonen. 2018. Post-specialisation: Retrofitting vectors of words unseen in lexical resources. In *Proceedings of NAACL-HLT 2018, New Orleans, Louisiana, June 1 - 6, 2018*, pages 516–527.

Xiangyu Wang and Chengqing Zong. 2021. Distributed representations of emotion categories in emotion space. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2364–2375.

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

Zhe Ye, Fang Li, and Timothy Baldwin. 2018. Encoding sentiment information into word vectors for sentiment analysis. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 997–1007.

Liang-Chih Yu, Jin Wang, K. Robert Lai, and Xuejie Zhang. 2017. Refining word embeddings for sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.*, pages 534 –539.