Disentangling Latent Emotions of Word Embeddings on Complex Emotional Narratives

Zhengxuan Wu[1][0000−0001−5581−8908] and Yueyi Jiang[2][0000−0002−5311−4648]

[1] Stanford University, Stanford CA 94085, USA
wuzhengx@stanford.edu
[2] University of California San Diego, La Jolla CA 92093, USA
yujiang@ucsd.edu

Abstract. Word embedding models such as GloVe are widely used in natural language processing (NLP) research to convert words into vectors. Here, we provide a preliminary guide to probe latent emotions in text through GloVe word vectors. First, we trained a neural network model to predict continuous emotion valence ratings by taking linguistic inputs from Stanford Emotional Narratives Dataset (SEND). After interpreting the weights in the model, we found that only a few dimensions of the word vectors contributed to expressing emotions in text, and words were clustered on the basis of their emotional polarities. Furthermore, we performed a linear transformation that projected high dimensional embedded vectors into an *emotion space*. Based on NRC Emotion Lexicon (EmoLex), we visualized the entanglement of emotions in the lexicon by using both projected and raw GloVe word vectors. We showed that, in the proposed *emotion space*, we were able to better disentangle emotions than using raw GloVe vectors alone. In addition, we found that the sum vectors of different pairs of emotion words successfully captured expressed human feelings in the EmoLex. For example, the sum of two embedded word vectors expressing *Joy* and *Trust* which express *Love* shared high similarity (similarity score .62) with the embedded vector expressing *Optimism*. On the contrary, this sum vector was dissimilar (similarity score -.19) with the the embedded vector expressing *Remorse*. In this paper, we argue that through the proposed *emotion space*, arithmetic of emotions is preserved in the word vectors. The affective representation uncovered in emotion vector space could shed some light on how to help machines to disentangle emotion expressed in word embeddings.

Keywords: Word embeddings · Emotional semantics · Affective computing.

1 Introduction

Constructing human-friendly Artificial Intelligence (AI) is essential for humans as it will help us get the most benefits from AI systems [13]. Being able to detect emotions through language is the building block of such AI agents [11, 17]. One such way is through word embeddings - encoding words in vectors. Researchers have proposed various word embedding methods such as GloVe and Word2Vec [10, 15]. However, to date, understanding the expressed human emotions in text from word embeddings by an

agent remains a challenging problem, as word embedding based models are generally missing the direct interpretations of the word vectors [1, 5].

In this study, we provide different strategies for interpreting the emotional semantics of words through word embeddings. We visualize word clusters by projecting word vectors into 2-dimensional space where embedding vectors are clustered by their emotional polarities. Additionally, based on the weights of a pretrained neural network model, we are able to project words into an *emotion space*. we show that the arithmetic of emotions holds, which is consistent with the principle introduced by Plutchik [16]. An example is as follows:

$$v_{\text{Love}} = v_{\text{Joy}} + v_{\text{Trust}} \tag{1}$$

We also show that words with opposite emotion valence separated in the *emotion space*. Rather than relying on dictionaries or hidden layers of neural networks, we provide a preliminary method of probing emotion entanglements in word vectors, making one initial step in exploring the latent emotions in word embeddings from modeling emotions in complex narratives.

2 Related Works

Word embeddings are widely applied in sentiment analysis with neural network models [2, 3, 19]. However, these models often lack clear interpretations of word vectors [7]. To date, only few studies have probed the semantics of emotions from word embeddings [8, 18]. These studies attempted to interpret natural language models through visualization of word vectors and hidden layers of the models. For example, researchers have visualized the hidden layers' representation of word vectors in 2-dimensional space in which words with similar meanings are clustered together [4, 6]. Additionally, other methods have focused on visualizing the hidden layers of neural network models using gradients and weights inferred from the models [7, 8, 18].

In this study, we aim to provide a systematic way of identifying emotions directly from text. Using embedded vectors, our method is different from the existing research that has focused on deriving latent semantic information from hidden layers of the neural network models [7, 8]. Throughout the paper, we provide preliminary evidence for detecting emotions in word vectors through word embeddings specifically in emotional expressions.

3 Dataset

In this paper, we used Stanford Emotional Narratives Dataset (SEND) as our dataset. SEND is comprised of transcripts of video recordings in which participants shared emotional stories, and it has been well explored in computational models of emotion [12, 19]. In each transcript, timestamps were generated for every word based on force-alignments[3] of audio inputs, and continuous emotional valence ratings were col-

[3] https://github.com/ucbvislab/p2fa-vislab

lected by annotators[4]. These ratings serve as the target variable in our model, which were scaled between [-1,1] and sampled every 0.5s.

The dataset includes 193 transcripts that last on average 2 mins 15 secs, for a total duration of 7 hrs and 15 mins. We divided these transcripts into a **Train set** (60% of the dataset, 117 videos, 38 targets, 4 hrs 26 mins long), a **Validation set** (20%, 38 videos, 27 targets, 1 hr 23 mins long) and a **Test set** (20%, 38 videos, 27 targets, 1 hr 26 mins long).

4 Autoregressive Model

To interpret the word vectors, we trained an autoregressive linear model to predict emotional valence ratings. We first used 300-dimensional GloVe word vectors which were pre-trained on wikipages [15]. Then, we used `interp` function in `numpy` package to assign each word a valence rating by linearly interpolating the original ratings using the timestamps for each word.

By concatenating the word vector $v_t \in \mathbb{R}^{e\times 1}$ where e is 300 for GloVe from the current time point (at time t) with the hidden state vector $h_{t-\tau} \in \mathbb{R}^{300\times 1}$ where h is 300 from last time point (at time $t-\tau$), we produced a hidden vector $h_t \in \mathbb{R}^{600\times 1}$ for the current time point. The hidden vector was then passed into a linear layer with bias to produce a output vector $o_t \in \mathbb{R}^{300\times 1}$. Subsequently, the output vector was passed into a linear layer which produced a single pseudo-rating prediction u_t for the current time point. The final rating prediction r_t was produced by a self-learned linear filter by taking $r_{t-\tau}$ into account:

$$h_t = \text{Concat}(h_{t-\tau}, v_t) \tag{2}$$

$$o_t = [\mathbf{W}_\text{h}, \mathbf{W}_\text{v}]h_t + \mathbf{b}_\text{h} \tag{3}$$

$$u_t = \mathbf{W}_\text{o}o_t + \mathbf{b}_\text{o} \tag{4}$$

$$r_t = \sigma r_{t-\tau} + (1-\sigma)u_t \tag{5}$$

with weight matrices $\mathbf{W}_\text{h}, \mathbf{W}_\text{v} \in \mathbb{R}^{300\times 300}$, $\mathbf{W}_\text{o} \in \mathbb{R}^{300\times 1}$ and bias vectors $\mathbf{b}_\text{h} \in \mathbb{R}^{600\times 300}$, $\mathbf{b}_\text{o} \in \mathbb{R}^{300\times 1}$. We used σ to denote the weight on previous rating prediction.

5 Model Evaluation

Before interpreting the model results, we ensured that the optimal performance was achieved. Similar to the evaluation metric used in a previous study on SEND [19], Concordance Correlation Coefficient (CCC, as defined by [9]) was evaluated. Specifically, we compared model performance on the **Validation set** and **Test set** with the human benchmark provided by SEND. Our model achieved a CCC of $.37 \pm .11$ on the **Validation set** and $.35 \pm .15$ on the **Test set**, comparing to human's performance of a CCC of $.47 \pm .12$ on the **Validation set** and $.46 \pm .14$ on the **Test set**. In our final model, the

[4] We have 25 ratings per transcript from annotators. The target variable is the average collected ratings.

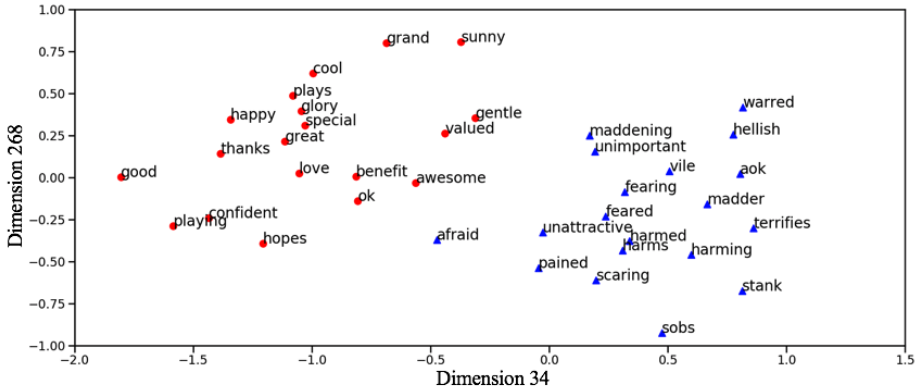

Fig. 1: Visualization of word clusters by their emotional polarities (i.e., positive or negative). Positive words are represented as red circles, while negative words are shown in blue triangles.

learnable parameter σ is 0.84, indicating that the current prediction at time t is primarily dependent on the previous state at time $t-1$.

6 Experiment And Results

6.1 Pure On Weights

By using the weights from two linear layers, each dimension was assigned a score to quantify its contribution to emotion valence in words (Alg.1). Higher absolute weights are associated with larger gradient changes in outputs, indicating higher importance in emotion expression in a given dimension.

Algorithm 1: Scoring algorithm for ranking important dimensions of GloVe vectors in emotion expression

Result: Scores for each dimension

dim_scores = {};
for $i \in \{1, ..., 300\}$ **do**
    _score = 0.0;
    for $w_j \in \boldsymbol{W}_{v_i}$ **do**
        _score += $w_j \cdot \mathbf{W}_{o_i}$;
    end
    dim_scores[i] = abs(_score);
end

Based on the scores, we first produced a heatmap (Fig.2.a) for all 300 dimensions in GloVe vectors to visualize the importance of each dimension. In addition, we plotted

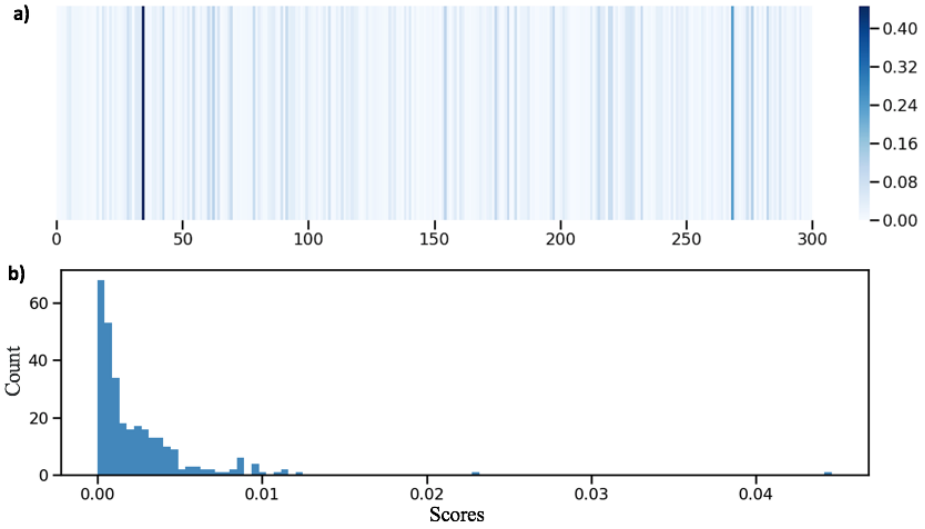

Fig. 2: Illustration of calculated scores on each dimension of the 300-dimensional GloVe vectors. (a) Heatmap of scores for each dimension in GloVe vectors. (b) Distribution of scores across all dimensions.

the distribution of scores (Fig.2.b). We found that a large portion of dimensions were non-expressive in emotion valence. Specifically, we discovered that the 34^{th} dimension of the GloVe vectors is the most important dimension in expressing emotions in word vectors.

6.2 2-d Emotion Visualization

Based on the scores for all dimensions, we picked out the top 2 dimensions and visualized the clustering effect of words. We used out-of-sample words from LIWC 2007 [14] from which we selected top 19 words ranked by their gradients of forward propagation for positive and negative polarities, respectively. Figure 1 shows that words with positive meaning are well separated from words with negative meaning in this space. Subsequently, we showed that the dimensions picked by our scoring algorithm (Alg.1) could be used to separate words into clusters that represent two emotion polarities, positive and negative emotions.

6.3 Entanglement Of Emotions

EmoLex has eight categories of word groups by emotions: *Joy*, *Trust*, *Anticipation*, *Surprise*, *Fear*, *Anger*, *Disgust* and *Sadness*. To show the entanglements of emotions in GloVe word vectors, we first randomly selected word pairs from any two distinct emotions of the eight emotion categories. We then exhaustively calculated the average

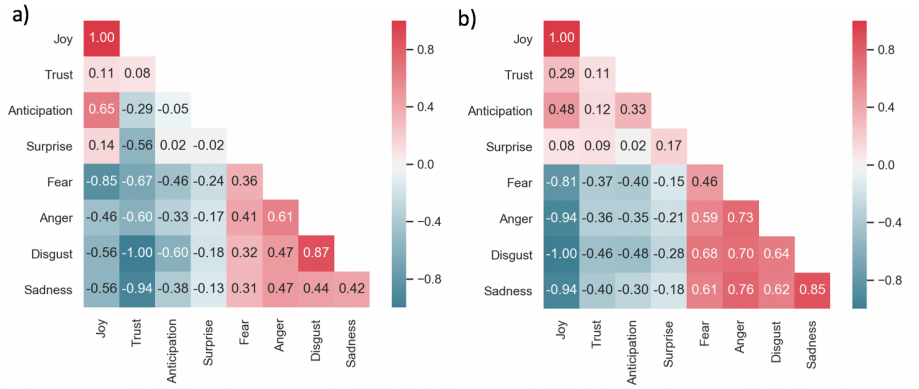

Fig. 3: Heatmaps of cosine similarities scores between words with paired emotions. (a) is produced by using the raw GloVe vector. (b) is produced by using the projected GloVe vectors.

cosine similarity scores between these word pairs and produced heatmaps with 8×8 similarity scores by

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{6}$$

To investigate if words are better clustered by emotional polarity in the proposed *emotion space*, we projected word vectors to this space by calculating element-wise multiplication of the weight $\mathbf{W}_{\text{v}}$ from our model and raw word vectors. Based on the heatmap (Fig. 3), we found that word vectors in the *emotion space* were better clustered by emotional polarity than raw GloVe vectors. For example, using the raw GloVe vectors, words expressing *Anticipation* had a low similarity score (-.29) with *Trust*, even though both words are associated with positive valence. However, after we projected the words into the *emotion space*, the similarity scores drastically increased. Thus, the results suggest that the projected GloVe vectors may provide better interpretations of emotion expressions in words.

6.4 Arithmetic Of Emotions

In this section, we demonstrate that with the linear projection matrix $\mathbf{W}_{\text{v}}$ from the pre-trained model, word vectors is transformed into a *emotion space* where the arithmetic of emotions (Tab. 1) is better represented compared to using raw GloVe vectors. According to Plutchiks wheel of emotions [16], feelings are combinations of two emotions. For example, *Love* is a combination of *Joy* and *Trust*. We want to examine if this arithmetic of emotions is preserved with word embeddings, which means whether the sum

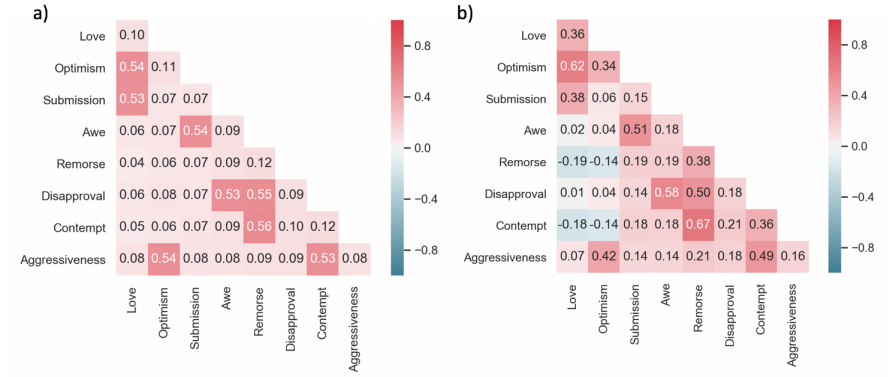

Fig. 4: Heatmaps of cosine similarities scores between words with paired feelings. (a) is produced by using the raw GloVe vector. (b) is produced by using the projected GloVe vectors.

Feelings	**Emotions**	**Opposite**	**Emotions**
Love	Joy + Trust	Remorse	Sadness + Disgust
Optimism	Anticipation + Joy	Disapproval	Surprise + Sadness
Submission	Trust + Fear	Contempt	Disgust + Anger
Awe	Fear + Surprise	Aggressiveness	Anger + Anticipation

Table 1: Taxonomy of feelings and arithmetic of emotions. The **Opposite** column represents a list of opposite feelings of the first column which is based on Plutchiks wheel of emotions [16].

of word vectors expressing *Joy* and *Trust* has a high similarity score with the word vector of *Love* . First, we formulated vectors representing the feeling that EmoLex is missing. For instance, we randomly paired words expressing *Joy* and *Trust* and added up word vectors from each pair to formulate a groups of vectors representing the feeling *Love*. Then, similar to the previous analysis, we calculated the average similarity scores between any two pairs of feelings using the generated word vectors. Ideally, the similarity score between two opposite feelings should be low whereas the score between two similar feelings should be high.

Based on our heatmap (Fig. 4), we found that arithmetic of emotions was not well preserved with raw GloVe vectors given the fact that the vectors of feelings had extremely low similarity scores with themselves. For example, the similarity scores between *Love* and itself is only .10. Meanwhile, opposite feelings had higher similarity scores than expected. In the *emotion space*, the distribution of similarity scores were more systematic. For example, the similarity scores between *Love* and itself increased to .36 whereas the similarity scores between two opposite feelings *Love* and *Remorse* decreased to -.19.

7 Conclusion

In the present study, we demonstrate that word embeddings with GloVe preserve latent emotions in text. By ranking weights across dimensions of the GloVe vectors, we show that majorities of dimensions are not associated with emotion representations in words. Additionally, from the top two dimensions ranked by importance, we demonstrate that words can be clustered by emotional polarity (Fig.1).

Using a projection matrix to transform the original GloVe vectors, we find that in the proposed *emotion space*, arithmetic of emotions is better represented compared to using the raw vectors alone. By comparing the similarities across vectors, we demonstrate that words with opposite emotional meanings are well separated in the *emotion space*. Meanwhile, we show that arithmetic of emotions is a good proxy of human feelings in the proposed space, consistent with the Plutchik's theory of emotions. Our preliminary exploration shed some lights on modeling the inter-relations in different emotion categories through word embeddings, and encourages more refined research in probing emotion expressions in other types of word embeddings.

References

1. Bordes, A., Weston, J., Usunier, N.: Open question answering with weakly supervised embedding models. In: Joint European conference on machine learning and knowledge discovery in databases. pp. 165–180. Springer (2014)
2. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2625–2634 (2015)
3. Ebrahimi Kahou, S., Michalski, V., Konda, K., Memisevic, R., Pal, C.: Recurrent neural networks for emotion recognition in video. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. pp. 467–474. ACM (2015)
4. Faruqui, M., Dyer, C.: Improving vector space word representations using multilingual correlation. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. pp. 462–471 (2014)
5. Gu, J., Bradbury, J., Xiong, C., Li, V.O., Socher, R.: Non-autoregressive neural machine translation. arXiv preprint arXiv:1711.02281 (2017)
6. Ji, Y., Eisenstein, J.: Representation learning for text-level discourse parsing. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). vol. 1, pp. 13–24 (2014)
7. Li, J., Chen, X., Hovy, E., Jurafsky, D.: Visualizing and understanding neural models in nlp. arXiv preprint arXiv:1506.01066 (2015)
8. Li, M., Lu, Q., Long, Y., Gui, L.: Inferring affective meanings of words from word embedding. IEEE Transactions on Affective Computing **8**(4), 443–456 (2017)
9. Lin, L.I.K.: A concordance correlation coefficient to evaluate reproducibility. Biometrics pp. 255–268 (1989)
10. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
11. Morelli, S.A., Ong, D.C., Makati, R., Jackson, M.O., Zaki, J.: Empathy and well-being correlate with centrality in different social networks. Proceedings of the National Academy of Sciences **114**(37), 9843–9847 (2017)

12. Ong, D.C., Wu, Z., Zhi-Xuan, T., Reddan, M., Kahhale, I., Mattek, A., Zaki, J.: Modeling emotion in complex stories: the Stanford Emotional Narratives Dataset (Invited Revision to Journal)
13. Ong, D.C., Zaki, J., Goodman, N.D.: Affective cognition: Exploring lay theories of emotion. Cognition **143**, 141–162 (2015)
14. Pennebaker, J.W., Francis, M.E., Booth, R.J.: Linguistic inquiry and word count: Liwc 2001. Mahway: Lawrence Erlbaum Associates **71**(2001), 2001 (2001)
15. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
16. Plutchik, R.: A general psychoevolutionary theory of emotion. In: Theories of emotion, pp. 3–33. Elsevier (1980)
17. Preston, S.D., De Waal, F.B.: Empathy: Its ultimate and proximate bases. Behavioral and Brain Sciences **25**(1), 1–20 (2002)
18. Seyeditabari, A., Zadrozny, W.: Can word embeddings help find latent emotions in text? preliminary results. In: The Thirtieth International Flairs Conference (2017)
19. Wu, Z., Zhang, X., Zhi-Xuan, T., Zaki, J., Ong, D.C.: Attending to emotional narratives. IEEE Affective Computing and Intelligent Interaction (ACII) (2019)