Aff2Vec: Affect–Enriched Distributional Word Representations

Sopan Khosla, Niyati Chhaya, and Kushal Chawla
Big Data Experience Lab
Adobe Research
Bangalore, India
{skhosla,nchhaya,kchawla}@adobe.com

Abstract

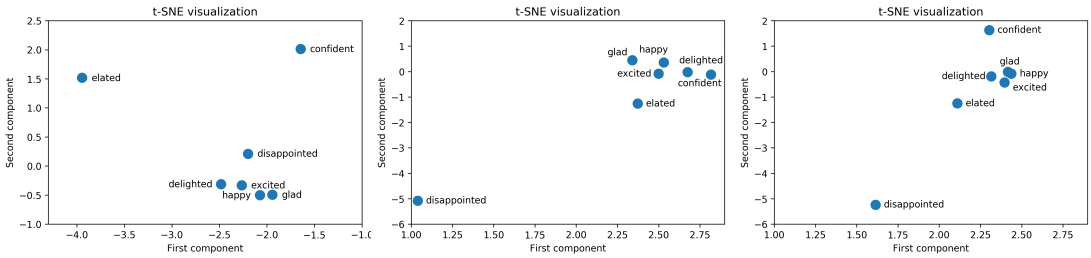
Human communication includes information, opinions and reactions. Reactions are often captured by the affective-messages in written as well as verbal communications. While there has been work in affect modeling and to some extent affective content generation, the area of affective word distributions is not well studied. Synsets and lexica capture semantic relationships across words. These models, however, lack in encoding affective or emotional word interpretations. Our proposed model, Aff2Vec, provides a method for enriched word embeddings that are representative of affective interpretations of words. Aff2Vec outperforms the state-of-the-art in intrinsic word-similarity tasks. Further, the use of Aff2Vec representations outperforms baseline embeddings in downstream natural language understanding tasks including sentiment analysis, personality detection, and frustration prediction.

1 Introduction

Affect refers to the experience of a feeling or an emotion (Scherer et al., 2010; Picard, 1997). This definition includes emotions, sentiments, personality, and moods. The importance of affect analysis in human communication and interactions has been discussed by Picard (1997). Historically, affective computing has focused on studying human communication and reactions through multi-modal data gathered via various sensors. The study of human affect from text and other published content is an important topic in language understanding. Word correlation with social and psychological processes is discussed by Pennebaker (2011). Preotiuc-Pietro et al. (2017) studied personality and psycho-demographic preferences through Facebook and Twitter content. Sentiment analysis in Twitter, with a detailed discussion on human affect (Rosenthal et al., 2017) and affect analysis in poetry (Kao and Jurafsky, 2012) have also been explored. Human communication not only contains semantic and syntactic information but also reflects the psychological and emotional states. Examples include the use of opinion and emotion words (Ghosh et al., 2017). The analysis of affect in interpersonal communication such as emails, chats, and longer articles is necessary for various applications including the study of consumer behavior and psychology, understanding audiences and opinions in computational social science, and more recently for dialogue systems and conversational agents. This is an open research space today.

Traditional natural language understanding systems rely on statistical language modeling and semantic word distributions such as WORDNET (Miller, 1995) to understand relationships across different words. There has been a resurgence of research efforts towards creating word distributions that capture multi-dimensional word semantics (Mikolov et al., 2013a; Pennington et al., 2014). Sedoc et al. (2017b) introduce the notion of affect features in word distributions but their approach is limited to creating enriched representations and no comments on the utility of the new word distribution is presented. Beyond word-semantics, deep learning research in natural language understanding is focused towards sentence representations using encoder-decoder models (Ahn et al., 2016), integration of symbolic knowledge to language models (Vinyals et al., 2015), and some recent works in augmenting neural language modeling with affective information to emotive text generation (Ghosh et al., 2017). These works, however, do

Proceedings of the 27th International Conference on Computational Linguistics, pages 2204–2218
Santa Fe, New Mexico, USA, August 20-26, 2018.



(a) GloVe (b) GloVe⊕Affect (c) GloVe + Counterfitting ⊕ Affect

Figure 1: t-SNE for significant affect words: The graphs show the distribution of sample words from Sedoc et al (2017b). The variance in the visualization illustrates the perturbation introduced by distributional schemes discussed in this paper. Vanilla GloVe embeddings show 'disappointed' near 'delighted', while these are separated in the ⊕Affect representations.

not introduce distributional affective word representations that not only reflect affective content but are also superior for related downstream natural language tasks such as sentiment analysis and personality detection.

We introduce Aff2Vec, affect-enriched word distributions trained on lexical resources coupled with semantic word distributions. Aff2Vec captures opinions and affect information in the representation using post-processing approaches. Figure 1 illustrates how Aff2Vec captures affective relationships using a t-SNE visualization of the word space. Even though Aff2Vec is trained on the Valence-Arousal-Dominance dimensions, our approach is generalizable to any other affect spaces. Our experiments show that Aff2Vec outperforms vanilla embedding spaces on both intrinsic word-similarity tasks as well as extrinsic natural language applications. The main contributions of this paper include:

- **Aff2Vec:** Affect-enriched word representations using post-processing techniques. We show that Aff2Vec outperforms the state-of-the-art in both intrinsic word similarity metrics as well as downstream natural language tasks including Sentiment analysis, Personality detection, and Frustration detection in interpersonal communication.

- **ENRON-FFP Dataset:** We introduce the ENRON-FFP Email dataset with Frustration, Formality, and Politeness tags gathered using a crowd-sourced human perception study.

The remainder of the paper is organized as follows. The prior art for enriched word distributions is discussed in Section 2. Aff2Vec is introduced in Section 3. We present a crowd-sourcing study for the ENRON-FFP Dataset in Section 4 and Section 5 discusses the experimental setup. Section 6 presents the evaluation of Aff2Vec for various intrinsic and extrinsic tasks. A discussion on the distributional word representations is presented in Section 7 before the conclusion in Section 8.

2 Related Work

The use of lexical semantic information (lexical resources) to improve distributional representations is recent. Methods like (Yu and Dredze, 2014; Xu et al., 2014; Bian et al., 2014; Kiela et al., 2015) achieve improved representations by using word similarity and relational knowledge to modify the prior or add a regularization term. We call such methods 'pre-training methods', as they alter the training process for word representations. Such methods require a change in the loss function while training the embeddings, hence are computationally expensive.

The other set of word distribution enhancements are done post-training. These methods aim to include external information using normalizations and modifications to the vanilla word distributions. Methods such as Retrofitting (Faruqui et al., 2015), which try to drag similar words closer together (where notion of similarity is taken from word relation knowledge found in semantic lexica (e.g. WordNet)) fall in this category. Counterfitting (Mrkšic et al., 2016) on the other hand, initiates from SimLex-999 tuned

embeddings, injects antonym and synonym constraints to improve word representations. This paper introduces post-training techniques on vanilla, retrofitted and counterfitted embeddings to include affective information in the distributions. Our work falls in the post-training category, hence no direct comparison with the pre-trained approaches is presented in this paper.

Recent work has explored approaches to adapt general-purpose lexica for specific contexts and affects. Studies have recognized the limited applicability of general purpose lexica such as ANEW (Bradley and Lang, 1999) to identify affect in verbs and adverbs, as they focus heavily on adjectives. Recognizing that general-purpose lexica often detect sentiment which is incongruous with context, Ribeiro et al. (2016) proposed a sentiment-damping method which utilizes the average sentiment strength over a document to damp any abnormality in the derived sentiment strength. Similarly, Blitzer et al. (2007) argued that words like 'predictable' induced a negative connotation to book reviews, while 'must-read' implied a highly positive sentiment. This paper doesn't focus on building yet another affect lexicon but studies the consequences of including affect information in distributional word representations that aim at defining relational relationships across all words in large contexts and vocabularies.

Automatic expansion of affect rating has been approached with the intuition that words closer in the distributional space would have similar ratings (Recchia and Louwerse, 2015; Palogiannidi et al., 2015; Vankrunkelsven et al., 2015; Köper and Im Walde, 2016). Recent work by Sedoc et al. (2017b) uses Signed Spectral Clustering to differentiate between words which are contextually similar but display opposite affect. Wang et al. (2016) use a graph-based method inspired by label propagation. While our approach follows the nature of the task defined in Sedoc et al. (2017b), we propose a generalized method to enrich content with affective information. Instead of only focusing on distinguishing the polarities, our method incorporates both semantic and affect information. Hence, creating embeddings that can also be used for semantic similarity tasks. Note that Sedoc et al. do not include any semantic information in their modeling.

3 Aff2Vec: Affect–enriched Word Distributions

Aff2Vec aims at incorporating affective information in word representations. We leverage the Warriner's lexicon (Warriner et al., 2013) in the Valence-Arousal-Dominance space for this work. The proposed work is generalizable to other affect spaces (Refer Appendix A for experiments with different dimensions.). This section presents two approaches for affect–enrichment of word distributions.

Warriner's lexicon: We use the Warriner's lexicon (Warriner et al., 2013) in this work. This is a affect lexicon with $13,915$ English words. It contains real-valued scores for valence, arousal, and dominance (VAD) on a scale of $1-9$ each. 1, 5, and 9 correspond to the low, moderate (i.e. neutral), and high values for each dimension respectively. For out-of-dictionary words, such as stop words or proper nouns, we assume a neutral affect vector $\vec{a} = [5, 5, 5]$.

3.1 Affect-APPEND ($\oplus$ Affect)

Consider word embeddings W, the aim is to introduce affective information to this space using the affect embedding space, A. The word vectors W, each with dimension D, are concatenated with affect vectors A with dimension F, thus resulting in a $D + F$ dimensional enriched representation. The process for this concatenation is described here:

1. Normalize word vector W and affect vector A using their L2-Norms (Equation 1). This reduces the individual vectors to unit-length.

$$x_i = \frac{x_i}{\sqrt{\sum_{k=1}^{D} x_{ik}^2}} \ \forall x_i \in W, \quad a_i = \frac{a_i}{\sqrt{\sum_{k=1}^{F} a_{ik}^2}} \ \forall a_i \in A \tag{1}$$

2. Concatenate the regularized word vectors x_i with regularized affect vectors a_i.

$$WA(w) = W(w) \oplus A(w) \tag{2}$$

3. Standardize (variance 1 , mean 0) the $D + F$ dimensional embeddings to achieve standard normal distribution.

$$y_i = \frac{y_i - \mu}{\sigma} \quad \forall y_i \in WA \tag{3}$$

where, μ and σ represent the mean and standard deviation respectively.

4. The enriched space WA is then reduced to original D dimensional vector. We use Principal Component Analysis for the dimensionality reduction.

3.2 Affect-STRENGTH

In this approach, the strength in the antonym-synonym relationships of the words is incorporated to the word distribution space. Hence, we leverage the retrofitting algorithm (Faruqui et al., 2015) as shown below. [1]

Retrofitting: Let $V = \{w_1, w_2, w_3, ..., w_n\}$ be a vocabulary and Ω be an ontology which encodes semantic relations between words present in V (e.g. WORDNET). This ontology Ω is represented as an undirected graph (V, E) with words as vertices and (w_i, w_j) as edges indicating the semantic relationship of interest. Each word $w_i \in V$ is represented as a vector representation $\hat{q}_i \in R^d$ learnt using a data-driven approach (e.g. Word2Vec or GloVe) where d is the length of the word vectors.

Let $\hat{Q}$ be the matrix collection of these vector representations. The objective is to learn the matrix $Q = (q_1, ..., q_n)$ such that the word vectors (q_i) are both close to their counterparts in $\hat{Q}$ and to adjacent vertices in Ω. The distance between a pair of vectors is defined to be Euclidean, hence the objective function for minimization is

$$\Psi(Q) = \sum_{i=1}^{n} \left[\alpha_i \|q_i - \hat{q}_i\|^2 + \sum_{(i,j)\in E} \beta_{ij} \|q_i - q_j\|^2 \right] \tag{4}$$

where, α and β are hyper parameters and control the relative strengths of the two associations. Ψ is a convex function in Q and its global optimal solution can be found by using an iterative update method. By setting $\frac{\partial \Psi(Q)}{\partial q_i} = 0$, the online updates are as follows:

$$q_i = \frac{\sum_{j:(i,j)\in E} \beta_{ij} q_j + \alpha_i \hat{q}_i}{\sum_{j:(i,j)\in E} \beta_{ij} + \alpha_i} \tag{5}$$

We propose two ways to modify β_{ij} in equation 4 in order to incorporate affective strength in the edge weights connecting two retrofitted vectors to each other.

Affect-cStrength ($*$ cStrength): In this approach, the affective strength is considered as a function of all F affect dimensions.

$$S(w_i, w_j) = 1 - \frac{\|a_i - a_j\|}{\sqrt{\sum_{f=1}^{F} max\_dist_f^2}} \tag{6}$$

where, a_i and a_j are F dimensional vectors in A and max_dist_f is defined as the maximum possible distance between two vectors in f^{th} dimension ($= 9.0 - 1.0 = 8.0$ for VAD dimensions).

Affect-iStrength ($*$ iStrength): Here, each dimension is treated individually. For every dimension f in A, we add an edge between neighbors in the Ontology Ω where the strength of that edge is given by $S_f(w_i, w_j)$:

$$S_f(w_i, w_j) = 1 - \frac{|a_{if} - a_{jf}|}{max\_dist_f}, \quad S(w_i, w_j) = \sum_{f=1}^{F} S_f(w_i, w_j) \tag{7}$$

β_{ij} from equation 5 is normalized with this strength function as $\beta_{ij} = \beta_{ij} * S(w_i, w_j)$, where $S(w_i, w_j)$ is defined by either Affect-cStrength or Affect-iStrength.

[1] https://github.com/mfaruqui/retrofitting

4 Dataset: ENRON-FFP

Table 1: Enron-FFP Dataset Description

Property	Value
Total number of emails (Main Experiment)	960
Total number of emails (Pilot Experiment)	90
Min. sentences per email	1
Max. sentences per email	17
Average email size (no. of sentences)	4.22
Average number of words per email	77.5

Table 2: Datasets for Intrinsic Evaluation

Dataset	# Word-Pairs
Word Similarity (**WS**) (Finkelstein et al., 2001)	353
RG-65 (Rubenstein and Goodenough, 1965)	65
MEN (Bruni et al., 2012)	3000
Miller-Charles (**MC**) (Miller and Charles, 1991)	30
RW (Luong et al., 2013)	2034
SCWS (Huang et al., 2012)	2023
SimLex-999 (**SL**) (Hill et al., 2016)	999
SimVerb-3500 (**SV**) (Gerz et al., 2016)	3500

Table 3: Example emails with varying inter-annotator agreements.

Affect Dimension	Example	Annotations
Frustration: Low Agreement	See highlighted portion. We should throw this back at Davis next time he points the finger.	(-1, -1, 0, 0, -2, -2, 0, 0, -2, 0)
Frustration: High Agreement	Please see announcement below. Pilar, Linda, India and Deb, please forward to all of your people. Thanks in advance, adr	(0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
Formality: Low Agreement	I talked with the same reporters yesterday (with Palmer and Shapro). Any other information that you can supply Gary would be appreciated. Steve, did Gary A. get your original as the CAISO turns email? GAC	(0, 0, -1, 1, 1, 1, 0, -1, -2, -1)
Politeness: High Agreement	John, This looks fine from a legal perspective. Everything in it is either already in the public domain or otherwise non-proprietary. Kind regards, Dan	(1, 1, 1, 1, 1, 1, 1, 1, 2, 1)

We introduce an email dataset, a subset of the ENRON data (Cohen, 2009), with tags about interpersonal communication traits: Formality, Politeness, and Frustration. Along with the content of the emails, the dataset also provides user and network level information for email exchanges between Enron employees.

Human Perceptions and Definitions: *Tone* or affects such as frustration and politeness are highly subjective. In this work, we do not attempt to introduce or standardize an accurate definition for frustration (or formality and politeness). Instead, we assume that these are defined by human perception and each individual may differ in their understanding of these metrics. This approach of using untrained human judgments has been used in prior studies of pragmatics in text data (Pavlick and Tetreault, 2016; Danescu-Niculescu-Mizil et al., 2013) and is a recommended way of gathering gold-standard annotations (Sigley, 1997). The tagged data is then used to predict the formality, frustration, and politeness tags using Aff2Vec embeddings.

Dataset Annotation: We conducted a crowd-sourced experiment using Amazon's Mechanical Turk[2]. The analysis presented in this section is based on $1,050$ emails that were tagged across multiple experiments[3]. Table 1 provides the statistics of the annotated data. We follow the annotation protocol of the Likert Scale (Allen and Seaman, 2007) for all three dimensions. Each email is considered as a single data point and only the text in the email body is provided for tagging. Frustration is tagged on a 3 point scale with neutral being equated to 'not frustrated'; 'frustrated' and 'very frustrated' are marked with -1 and -2 respectively. Formality and politeness follow a 5 point scale from -2 to $+2$, where both extremes mark the higher degree of presence and absence of the respective dimension. Table 3 shows example emails from the dataset.

Inter-annotator Agreement: To measure whether an individual's intuition of the affect dimensions is consistent with other annotators' judgment, we use inter-class correlation[4] to quantify the

[2]https://www.mturk.com/mturk/welcome

[3]Link to the annotated ENRON-FFP dataset: https://bit.ly/2IAxPab

[4]We report the average raters absolute agreement (ICC1k) using the psych package in R.

Table 4: Intrinsic Evaluation: Word Similarity–We report the Spearman's correlation coefficient (ρ). The results show that Aff2Vec variants improve performance consistently.

Model	**Word Similarity**							
	SL	**SV**	**WS**	**RG**	**RW**	**SCWS**	**MC**	**MEN**
GloVe	0.41	0.28	0.74	0.77	0.54	0.64	0.80	0.80
$\oplus$ Affect	0.49	0.39	**0.77**	0.79	0.59	0.67	0.80	0.84
+ Retrofitting	0.53	0.37	0.73	0.81	0.52	0.66	0.82	0.82
+ Retrofitting $*$ cStrength	0.53	0.36	0.74	0.81	0.52	0.66	0.82	0.82
+ Retrofitting $*$ iStrength	0.56	0.38	0.64	0.80	0.44	0.62	0.80	0.78
+ Retrofitting $\oplus$ Affect	0.60	0.46	0.76	0.81	**0.61**	**0.69**	0.81	**0.85**
+ Counterfitting	0.58	0.47	0.65	0.80	0.56	0.61	0.78	0.77
+ Counterfitting $\oplus$ Affect	**0.62**	**0.53**	0.70	**0.84**	**0.61**	0.64	**0.84**	0.80
Word2Vec	0.45	0.36	0.70	0.76	0.59	0.67	0.80	0.78
$\oplus$ Affect	0.49	0.42	0.67	0.81	0.59	0.66	0.85	0.79
+ Retrofitting	0.55	0.45	**0.74**	0.82	**0.62**	**0.70**	0.83	0.80
+ Retrofitting $*$ cStrength	0.55	0.44	0.73	0.82	0.62	**0.70**	0.83	0.80
+ Retrofitting $*$ iStrength	0.58	0.47	0.71	0.83	0.57	0.69	0.85	0.80
+ Retrofitting $\oplus$ Affect	0.59	0.49	0.71	**0.84**	**0.62**	**0.70**	**0.86**	**0.82**
+ Counterfitting	0.56	0.51	0.66	0.75	0.61	0.64	0.75	0.73
+ Counterfitting $\oplus$ Affect	**0.60**	**0.54**	0.64	0.82	0.60	0.64	0.82	0.76
Paragram	0.69	0.54	**0.73**	0.78	0.59	0.68	0.80	0.78
$\oplus$ Affect	0.71	0.59	0.70	0.77	**0.60**	0.67	0.76	**0.79**
+ Retrofitting	0.68	0.55	**0.73**	0.79	0.59	0.68	0.81	0.78
+ Retrofitting $*$ cStrength	0.69	0.55	**0.73**	0.79	0.59	**0.69**	0.81	0.78
+ Retrofitting $*$ iStrength	0.68	0.56	0.71	0.80	0.58	0.68	**0.84**	0.77
+ Retrofitting $\oplus$ Affect	0.71	0.58	0.70	0.80	0.59	0.67	0.78	**0.79**
+ Counterfitting	0.74	0.63	0.69	**0.81**	**0.60**	0.66	0.82	0.74
+ Counterfitting $\oplus$ Affect	**0.75**	**0.66**	0.68	**0.81**	**0.60**	0.65	0.82	0.76

ordinal ratings. This measure accounts for the fact that we may have different groups of annotators for each data point. Each data point has 10 distinct annotations. Agreements reported are 0.506 ± 0.05(for 3 class), 0.73 ± 0.02(for 5 class), and 0.64 ± 0.03(for 5 class) for frustration, formality, and politeness respectively. The agreement measures are similar to those reported for other such psycholinguistic tagging tasks.

5 Experiments

Two sets of experiments are presented to evaluate Aff2Vec embeddings[5] - Intrinsic evaluation using word similarity tasks and extrinsic evaluation using multiple NLP applications. We focus on 3 vanilla word embeddings: GloVe (Pennington et al., 2014), Word2Vec-SkipGram[6] (Mikolov et al., 2013b) and Paragram-SL999 (Wieting et al., 2015); and their retrofitted (Faruqui et al., 2015) and counterfitted (Mrkšic et al., 2016) versions. The vocabulary and embeddings used in our experiments resonate with the experimental setup by Mrkšić et al.(2016) ($76,427$ words).

5.1 Intrinsic Evaluation

Word similarity is a standard task used to evaluate embeddings (Mrkšic et al., 2016; Faruqui et al., 2015; Bollegala et al., 2016). In this paper, we evaluate the embeddings on benchmark datasets given in Table 2. We report the Spearman's rank correlation coefficient between rankings produced by our model (based on cosine similarity of the pair of words) against the benchmark human rankings for each dataset.

5.2 Extrinsic Evaluation

Although intrinsic tasks are popular, performance of word embeddings on these benchmarks does not reflect directly into the downstream nlp tasks (Chiu et al., 2016). Gladkova and Drozd (2016) and

[5]Link to the Aff2Vec word embeddings: https://bit.ly/2HGohsO

[6]https://code.google.com/archive/p/word2vec/

Table 5: Extrinsic Evaluation: Results for FFP-Prediction, Personality Detection, Sentiment Analysis, and WASSA Emotional Intensity task for Aff2Vec variants for GloVe and Word2Vec embeddings. We report the Mean Squared Error (MSE) for FFP-Prediction, Accuracy (% ACC) for Personality Detection, and Sentiment Analysis (SA) and Person's ρ for the WASSA Emo-Int Task (EMO-INT)

Model	**FFP-Prediction**			**Personality Detection**					**SA**	**EMO-INT**			
	MSE ($X10^{-3}$)			**Acc.** (%)					**Acc.** (%)	**Pearson's** ρ ($X10^{-2}$)			
	FOR	**FRU**	**POL**	**EXT**	**NEU**	**AGR**	**CON**	**OPEN**	**DAN**	**ANG**	**FEA**	**JOY**	**SAD**
GloVe	27.59	32.40	21.89	**56.08**	55.25	56.06	**57.32**	59.14	83.1	70.98	71.19	65.85	73.30
$\oplus$ Affect	27.72	28.76	22.02	51.47	57.41	56.09	55.06	**62.08**	84.3	70.91	71.72	66.26	**73.58**
+ Retrofitting	27.44	29.35	21.75	55.79	59.67	55.59	56.89	59.67	82.7	72.10	71.86	**67.11**	73.14
+ Retrofitting $\oplus$ Affect	28.33	**27.91**	22.24	55.01	56.43	**57.48**	53.04	61.12	83.7	**72.38**	**72.53**	66.29	72.76
+ Counterfitting	**25.66**	29.20	22.90	55.11	58.32	55.41	53.89	60.36	84.2	70.45	68.95	65.27	72.63
+ Counterfitting $\oplus$ Affect	28.89	32.46	**21.64**	52.12	**60.03**	56.53	54.93	59.51	**84.4**	70.20	70.43	65.81	72.37
Word2Vec	25.86	27.88	21.56	**56.08**	58.19	56.59	55.18	61.41	83.3	68.86	71.24	65.23	72.60
$\oplus$ Affect	25.39	28.16	22.99	53.54	57.97	55.17	54.12	59.31	83.4	69.29	**71.92**	64.49	**72.63**
+ Retrofitting	27.81	29.05	21.85	54.33	56.65	**57.39**	54.65	60.03	82.5	70.12	71.42	**67.96**	72.02
+ Retrofitting $\oplus$ Affect	**25.08**	**27.08**	21.64	53.74	**59.61**	56.34	**56.93**	59.7	83.3	**70.65**	71.90	66.36	72.20
+ Counterfitting	28.28	27.12	22.95	54.55	57.61	57.09	54.1	58.5	83.3	68.64	70.13	63.36	70.67
+ Counterfitting $\oplus$ Affect	27.73	29.67	**21.52**	51.28	58.86	56.66	53.22	**61.62**	**83.5**	69.38	70.31	64.94	71.37
Baselines													
(Majumder et al., 2017)	–	–	–	**58.09**	59.38	56.71	57.30	**62.68**	–	–	–	–	–
ENRON Trainable	31.61	43.90	26.27	–	–	–	–	–	–	–	–	–	–
Re(Glove)(Yu et al., 2017)	–	–	–	–	–	–	–	–	82.2	–	–	–	–
Re(w2v)(Yu et al., 2017)	–	–	–	–	–	–	–	–	82.4	–	–	–	–

Batchkarov et al. (2016) suggest that intrinsic tasks should not be considered as gold standards but as a tool to improve the model. Therefore, we test the utility of Aff2Vec on 4 distinct natural language understanding tasks:

Affect Prediction (FFP-Prediction): The experiment is to predict the formality, politeness, and frustration in email. We introduce the ENRON-FFP dataset for this task in section 4. A basic CNN model is used for the prediction (Refer to Appendix B.4 for hyper-parameters and model details). The purpose of this experiment is to evaluate the quality of the embeddings and not necessarily the model architecture. The CNN is hence not optimized for this task. Embeddings trained on the ENRON dataset (ENRON-Trainable) are used as a baseline.

Personality Detection: This task is to predict human personality from text. The big five personality dimensions (Digman, 1990) are used for this experiment. The 5 personality dimensions include Extroversion (EXT), Neuroticism (NEU), Agreeableness (AGR), Conscientiousness (CON), and Openness (OPEN). Stream-of-consciousness essay dataset by Pennebaker et al. (1999) contains $2,468$ anonymous essays tagged with personality traits of the author. We use this dataset for the experiment. Majumder et al (2017) propose a CNN model for this prediction. We use their best results as baseline and report the performance of Aff2Vec on their default implementation[7].

Sentiment Analysis: The Stanford Sentiment Treebank (SST) (Socher et al., 2013) contains sentiment labels on sentences from movie reviews. This dataset in its binary form is split into $6,920$ training, 872 validation, and $1,821$ test set samples. We report the performance on a Deep Averaging Network (DAN)(Iyyer et al., 2015)[8] with default parameters on the SST dataset and compare against refined embeddings specifically created for sentiment analysis. Implementation by Yu et al (2017) is used for the refined embeddings[9].

Emotion Intensity Task (WASSA): WASSA shared task on emotion intensity (Mohammad and Bravo-Marquez, 2017) requires to determine the intensity of a particular emotion (anger, fear, joy, or

[7]https://github.com/SenticNet/personality-detection

[8]https://github.com/miyyer/dan

[9]Implementation provided by the authors is used for this experiment.

Table 6: Polarity-Noise@k (PN@10) and Granularity-Noise@k (GN@10) where $k = 10$ for GloVe and Word2Vec variants. Note that lower the number, better this qualitative metric.

Model	**PN@10** (%)			**GN@10** ($X10^{-2}$)		
	V	**A**	**D**	**V**	**A**	**D**
GloVe	23.21	22.15	27.07	83.91	79.19	74.19
⊕ Affect	**16.46**	19.65	**19.42**	**72.56**	**69.00**	64.02
+ Retrofitting	22.55	21.82	26.5	82.15	78.68	72.53
+ Retrofitting ∗ cStrength	22.07	21.63	26.14	80.85	78.12	71.86
+ Retrofitting ∗ iStrength	23.05	21.77	26.66	83.14	78.76	72.65
+ Retrofitting ⊕ Affect	19.68	**18.16**	22.88	73.45	71.56	66.55
+ Counterfitting	22.68	22.2	26.46	83.31	78.78	72.54
+ Counterfitting ⊕ Affect	16.75	19.99	19.99	73.89	69.55	**63.93**
Word2Vec	24.66	22.19	27.41	85.81	79.23	74.25
⊕ Affect	20.62	**17.83**	23.19	**74.78**	**71.64**	67.32
+ Retrofitting	23.75	22.25	26.94	84.65	79.36	73.00
+ Retrofitting ∗ cStrength	23.33	22.01	26.58	83.39	78.71	72.24
+ Retrofitting ∗ iStrength	23.90	22.30	27.13	85.34	79.46	73.12
+ Retrofitting ⊕ Affect	20.61	18.54	23.6	75.71	72.47	67.61
+ Counterfitting	23.47	22.48	26.72	84.62	79.14	72.29
+ Counterfitting ⊕ Affect	**20.34**	18.17	**23.01**	74.83	71.94	**66.62**
Paragram	25.16	22.55	28.05	88.34	80.73	75.49
⊕ Affect	20.81	21.29	23.45	81.83	75.27	69.79
+ Retrofitting	25.69	22.8	28.48	89.67	81.25	76.05
+ Retrofitting ∗ cStrength	25.46	22.64	28.22	89.06	80.95	75.58
+ Retrofitting ∗ iStrength	25.69	22.84	28.43	89.85	81.26	75.93
+ Retrofitting ⊕ Affect	23.38	**20.34**	25.99	83.17	76.51	71.83
+ Counterfitting	24.86	22.76	27.88	88.27	80.68	75.18
+ Counterfitting ⊕ Affect	**20.31**	21.50	**23.03**	**81.40**	**75.05**	**69.10**

sadness) in a tweet. This intensity score can be seen as an approximation of the emotion intensity of the author or as felt by the reader. We train a BiLSTM-CNN–based model for this regression task with embedding dimensions as features (Refer to Appendix B.3 for model details). Vanilla embeddings are used as a baseline for this experiment.

5.3 Qualitative Evaluation: Noise@k

Affect-enriched embeddings perform better as they move semantically similar but affectively dissimilar words away from each other in the vector space. We demonstrate this effect through two measures that capture noise in the neighborhood of a word.

Polarity-Noise@k (PN@k) (Yu et al., 2017) calculates the number of top k nearest neighbors of a word with opposite polarity for the affect dimension under consideration.
Granular-Noise@k (GN@k) captures the average difference between a word and its top k nearest neighbors for a particular affect dimension (f).

$$GN_i@k = \frac{\sum_{j \in kNN_i} |a_i f - a_j f|}{k} \tag{8}$$

where, a_i, a_j are F-dimensional vectors in A and kNN_i denotes the top k nearest neighbors of word i. This is done for each word in the affect lexicon.

6 Results

All experiments are compared against the vanilla word embeddings, embeddings with counterfitting, and embeddings with retrofitting.

Table A.1 summarizes the results of the **Intrinsic word–similarity tasks**. For the pre-trained word embeddings, Paragram-SL999 outperformed GloVe and Word2Vec on most metrics. Both retrofitting and counterfitting procedures show better or at par performance on all datasets except for WordSim-353. Addition of affect information to different versions of GloVe consistently improves performance

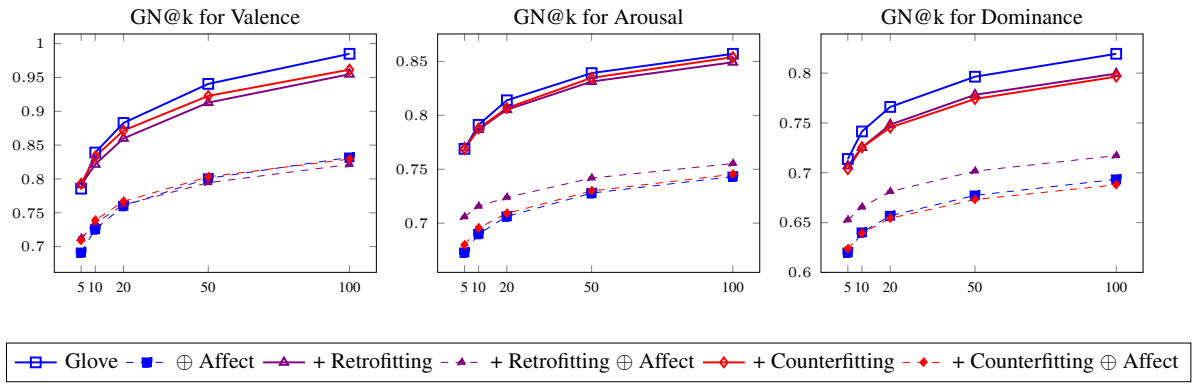


Figure 2: Variation of Granular Noise with different k values for GloVe and Affect-APPEND variants

whereas the only significant improvement for Paragram-SL999 variants is observed on the SimLex-999 and SimVerb-3500 datasets. To the best of our knowledge, $\rho = 0.74$ reported by Mrkšić et al. (2016) represents the current state-of-the-art for SimLex-999 and inclusion of affect information to these embeddings yields higher performance ($\rho = 0.75$). Similarly, for the SimVerb-3500 dataset, Paragram+Counterfitting⊕Affect embeddings beat the state-of-the-art scores[10]. Amongst Affect-APPEND and Affect-STRENGTH, Affect-APPEND out performs the rest in most cases for GloVe and Word2Vec. However, Affect-STRENGTH variations perform slightly better for the Paragram embeddings.

The results for the **Extrinsic tasks** are reported in Table 5. We report the performance for GloVe and Word2Vec with Affect-APPEND variants[11]. For FFP-Prediction, Affect-APPEND gives lowest Mean Squared Error for Frustration and Politeness. However, in the case of Formality, the counterfitting variant reports the lowest error. For the personality detection task, Affect-APPEND variants report best performance for NEU, AGR, and OPEN classes. For CON, Glove beats the best results in (Majumder et al., 2017). Evaluation against the Sentiment Analysis(SA) task shows that Affect-APPEND variants report highest accuracies. The final experiment reported here is the WASSA-EmoInt task. Affect-APPEND and retrofit variants out-perform the vanilla embeddings.

To summarize, the extrinsic evaluation supports the hypothesis that affect–enriched embeddings improve performance for all NLP tasks. Further, the word similarity metrics show that Aff2Vec is not specific to sentiment or affect-related tasks but is at par with accepted embedding quality metrics.

Qualitative Evaluation: Table 6 reports the average *Polarity-Noise@10* and *Granular-Noise@10* for GloVe, Word2Vec, and Paragram-SL999 variants. Note that lower the noise better the performance. Affect-APPEND reports the lowest noise for all cases. This shows that the introduction of affect dimensions in the word distributions intuitively captures psycholinguistic and in particular polarity properties in the vocabulary space. The rate of change of noise with varying k provides insights into (1) how similar are the embedding spaces and (2) how robust are the new representations to the noise i.e. how well is the affect captured in the new embeddings. Figure 2 shows the granular noise@k for valence, arousal, and sominance respectively. Noise@k for the Aff2Vec i.e. the Affect-APPEND variants, specifically, ⊕Affect and Couterfitting⊕Affect has lower noise even for a higher k. The growth rate for all variants is similar and reduces with an increase in the value of k. A similar behavior is observed for Polarity-Noise@k.

7 Discussion

Experiments give an empirical evaluation of the proposed embeddings, none of these provide an insight about the change in the distributional representations of the associated words. Semantic relationship

[10]Mentioned at `http://people.ds.cam.ac.uk/dsg40/simverb.html`

[11]Results for Paragram are reported in the supplement.

Table 7: Top-5 NN for 'Good' and 'Bad' for variants of GloVe, SentiWordNet and Aff2Vec

Model	Good	Bad
GloVe	[*great*, *nice*, *excellent*, *decent*, *bad*]	[*terrible*, *awful*, *horrible*, *wrong*, *thing*]
⊕ Affect	[*great*, *nice*, *excellent*, *decent*, *pretty*]	[*awful*, *terrible*, *horrible*, *wrong*, *crappy*]
+ Retrofitting	[*great*, *decent*, *nice*, *excellent*, *pretty*]	[*wrong*, *awful*, *terrible*, *horrible*, *nasty*]
+ Retrofitting ⊕ Affect	[*nice*, *great*, *decent*, *excellent*, *pretty*]	[*awful*, *wrong*, *nasty*, *terrible*, *horrible*]
+ Counterfitting	[*decent*, *nice*, *optimum*, *presentable*, *exemplary*]	[*rotten*, *shitty*, *horrid*, *naughty*, *lousy*]
+ Counterfitting ⊕ Affect	[*nice*, *decent*, *optimum*, *presentable*, *dignified*]	[*rotten*, *shitty*, *horrid*, *lousy*, *naughty*]
Senti-WordNet[12]	[*commodity*, *full*, *estimable*, *beneficial*, *adept*]	[*regretful*, *badly*]
Warriner's Lexicon	[*grandmother*, *healing*, *cheesecake*, *play*, *blissful*]	[*jittery*, *fuss*, *incessant*, *tramp*, *belligerent*]

capture the synonym like information. We study how the **neighborhood** of a certain word changes based on the different word distribution techniques used to create the corresponding representations. Table 7 shows the top five nearest neighbors based on the representations used. While SENTI-Wordnet represents synonyms more than affectively similar words, the affect–enriched embeddings provide a combination of both affective and semantic similarity. The variance in the ranking of words depicts how different schemes capture the intuition of word distributions. Such an analysis can be used to build automated natural language generation and text modification systems with varying objectives.

8 Conclusion

We present a novel, simple yet effective method to create affect–enriched word embeddings using affect and semantic lexica. The proposed embeddings outperform the state-of-the-art in benchmark intrinsic evaluations as well as extrinsic applications including sentiment analysis, personality detection, and affect prediction. We introduce a new human-annotated dataset with formality, politeness, and frustration tags on a subset of the publicly available ENRON email data. We are currently exploring the effect of dimension size on the performance of the enriched embeddings as well as the use of Aff2Vec for complex tasks such as text generation.

References

Sungjin Ahn, Heeyoul Choi, Tanel Pärnamaa, and Yoshua Bengio. 2016. A neural knowledge language model. *CoRR*, abs/1608.00318.

I Elaine Allen and Christopher A Seaman. 2007. Likert scales and data analyses. *Quality progress*, 40(7):64.

Miroslav Batchkarov, Thomas Kober, Jeremy Reffin, Julie Weeds, and David Weir. 2016. A critique of word similarity as a method for evaluating distributional semantic models.

Jiang Bian, Bin Gao, and Tie-Yan Liu. 2014. Knowledge-powered deep learning for word embedding. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 132–148. Springer.

John Blitzer, Mark Dredze, Fernando Pereira, et al. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*, volume 7, pages 440–447.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association of Computational Linguistics*, 5(1):135–146.

Danushka Bollegala, Mohammed Alsuhaibani, Takanori Maehara, and Ken-ichi Kawarabayashi. 2016. Joint word representation learning using a corpus and a semantic lexicon. In *AAAI*, pages 2690–2696.

Margaret M Bradley and Peter J Lang. 1999. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Technical report C-1, the center for research in psychophysiology, University of Florida.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 136–145. Association for Computational Linguistics.

[12]http://sentiwordnet.isti.cnr.it/

Billy Chiu, Anna Korhonen, and Sampo Pyysalo. 2016. Intrinsic evaluation of word vectors fails to predict extrinsic performance. In *Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP*, pages 1–6.

William W Cohen. 2009. Enron email dataset.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *51st Annual Meeting of the Association for Computational Linguistics*, pages 250–259. ACL.

John M Digman. 1990. Personality structure: Emergence of the five-factor model. *Annual review of psychology*, 41(1):417–440.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):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.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414. ACM.

Daniela Gerz, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. Simverb-3500: A large-scale evaluation set of verb similarity. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2173–2182.

Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2017. Affect-lm: A neural language model for customizable affective text generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 634–642. Association for Computational Linguistics.

Anna Gladkova and Aleksandr Drozd. 2016. Intrinsic evaluations of word embeddings: What can we do better? In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 36–42.

Felix Hill, Roi Reichart, and Anna Korhonen. 2016. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*.

Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 873–882. Association for Computational Linguistics.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1681–1691.

Justine Kao and Daniel Jurafsky. 2012. A computational analysis of style, affect, and imagery in contemporary poetry. In *CLfL@NAACL-HLT*.

Douwe Kiela, Felix Hill, and Stephen Clark. 2015. Specializing word embeddings for similarity or relatedness. In *EMNLP*, pages 2044–2048.

Maximilian Köper and Sabine Schulte Im Walde. 2016. Automatically generated affective norms of abstractness, arousal, imageability and valence for 350 000 german lemmas. In *LREC*.

Maximilian Köper, Evgeny Kim, and Roman Klinger. 2017. Ims at emoint-2017: emotion intensity prediction with affective norms, automatically extended resources and deep learning. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 50–57.

Thang Luong, Richard Socher, and Christopher D Manning. 2013. Better word representations with recursive neural networks for morphology. In *CoNLL*, pages 104–113.

Navonil Majumder, Soujanya Poria, Alexander Gelbukh, and Erik Cambria. 2017. Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems*, 32(2):74–79.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, pages 55–60.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

George A Miller and Walter G Charles. 1991. Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Saif Mohammad and Felipe Bravo-Marquez. 2017. Wassa-2017 shared task on emotion intensity. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 34–49. Association for Computational Linguistics.

Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2, pages 321–327.

Nikola Mrkšic, Diarmuid OSéaghdha, Blaise Thomson, Milica Gašic, 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*, pages 142–148.

Elisavet Palogiannidi, E Losif, Polychronis Koutsakis, and Alexandros Potamianos. 2015. Valence, arousal and dominance estimation for english, german, greek, portuguese and spanish lexica using semantic models.

Ellie Pavlick and Joel Tetreault. 2016. An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics*, 4:61–74.

James W Pennebaker and Laura A King. 1999. Linguistic styles: Language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296.

J.W. Pennebaker. 2011. *The Secret Life of Pronouns: What Our Words Say About Us*. Bloomsbury USA.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Rosalind W. Picard. 1997. *Affective Computing*. MIT Press, Cambridge, MA, USA.

Daniel Preotiuc-Pietro, Ye Liu, Daniel J. Hopkins, and Lyle Ungar. 2017. Personality Driven Differences in Paraphrase Preference. In *Proceedings of the Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, ACL.

Gabriel Recchia and Max M Louwerse. 2015. Reproducing affective norms with lexical co-occurrence statistics: Predicting valence, arousal, and dominance. *The Quarterly Journal of Experimental Psychology*, 68(8):1584–1598.

Filipe N Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabrício Benevenuto. 2016. Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5(1):1–29.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518.

Sascha Rothe, Sebastian Ebert, and Hinrich Schütze. 2016. Ultradense word embeddings by orthogonal transformation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 767–777.

Herbert Rubenstein and John B Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Klaus R. Scherer, Tanja Banziger, and Etienne Roesch. 2010. *A Blueprint for Affective Computing: A Sourcebook and Manual*. Oxford University Press, Inc., New York, NY, USA, 1st edition.

Joao Sedoc, Jean Gallier, Dean Foster, and Lyle Ungar. 2017a. Semantic word clusters using signed spectral clustering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 939–949, Vancouver, Canada, July. Association for Computational Linguistics.

Joao Sedoc, Daniel Preotiuc-Pietro, and Lyle Ungar. 2017b. Predicting Emotional Word Ratings using Distributional Representations and Signed Clustering. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, EACL.

Robert J Sigley. 1997. Text categories and where you can stick them: a crude formality index. *International Journal of Corpus Linguistics*, 2(2):199–237.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Hendrik Vankrunkelsven, Steven Verheyen, Simon De Deyne, and Gerrit Storms. 2015. Predicting lexical norms using a word association corpus. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, pages 2463–2468. Cognitive Science Society.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164.

Ivan Vulić, Nikola Mrkšić, Roi Reichart, Diarmuid Ó Séaghdha, Steve Young, and Anna Korhonen. 2017. Morph-fitting: Fine-tuning word vector spaces with simple language-specific rules. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 56–68.

Jin Wang, Liang-Chih Yu, K Robert Lai, and Xuejie Zhang. 2016. Community-based weighted graph model for valence-arousal prediction of affective words. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11):1957–1968.

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.

John Wieting, Mohit Bansal, Kevin Gimpel, Karen Livescu, and Dan Roth. 2015. From paraphrase database to compositional paraphrase model and back. *Transactions of the Association for Computational Linguistics*, 3:345–358.

Chang Xu, Yalong Bai, Jiang Bian, Bin Gao, Gang Wang, Xiaoguang Liu, and Tie-Yan Liu. 2014. Rc-net: A general framework for incorporating knowledge into word representations. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1219–1228. ACM.

Mo Yu and Mark Dredze. 2014. Improving lexical embeddings with semantic knowledge. In *ACL (2)*, pages 545–550.

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.

Appendix A Generalization of Affect and Emotion Dimensions

Apart from the Warriner's lexicon (Warriner et al., 2013) which is in the VAD space, we experimented with the NRC Affect Intensity lexicon by Mohammad (2013) and Ekman's Six emotions from IMS EmoInt Norms dataset[13](Köper et al., 2017; Ekman, 1992). The NRC lexicon has 4 dimensions whereas the Ekman has 6 dimensions. Table A.1 shows the intrinsic word similarity measures for both these affect spaces. The enriched embeddings reported here are based on Affect-APPEND. We report variants of the lexica with vanilla embeddings and their retrofitted and counterfitted versions. The intrinsic metrics improve over the vanilla embeddings with the Aff2Vec versions. Note that the improvement achieved here is slightly lower than those achieved with the VAD space.

This analysis supports the hypothesis that Aff2Vec is generalizable to other dimension spaces and not restricted to a specific affect distribution.

Table A.1: Intrinsic Evaluation: Word Similarity–We report the Spearman's correlation coefficient (ρ) against NRC as well as Ekman's dimensions. The results show that Aff2Vec variants improve performance consistently

Model	**Word Similarity**							
	SL	**SV**	**WS**	**RG**	**RW**	**SCWS**	**MC**	**MEN**
GloVe	0.41	0.28	0.74	0.77	0.54	0.64	0.80	0.80
$\oplus$ Ekman's Six	0.46	0.34	0.77	0.78	0.58	0.67	0.80	0.83
+ Retrofitting $\oplus$ Ekman's Six	0.57	0.42	**0.79**	0.83	**0.61**	**0.70**	**0.82**	**0.85**
+ Counterfitting $\oplus$ Ekman's Six	**0.61**	**0.50**	0.60	0.83	**0.61**	0.64	0.80	0.80
$\oplus$ NRC Affect	0.47	0.34	0.78	0.78	0.58	0.67	0.81	0.84
+ Retrofitting $\oplus$ NRC Affect	0.58	0.43	**0.79**	0.82	**0.61**	**0.70**	0.81	**0.85**
+ Counterfitting $\oplus$ NRC Affect	**0.61**	**0.50**	0.70	**0.84**	**0.61**	0.64	0.81	0.80
Word2Vec	0.45	0.36	0.70	0.76	0.59	0.67	0.80	0.78
$\oplus$ Ekman's Six	0.46	0.38	0.69	0.80	0.59	0.67	0.82	0.79
+ Retrofitting $\oplus$ Ekman's Six	0.57	0.45	**0.73**	**0.85**	0.61	0.70	**0.85**	**0.82**
+ Counterfitting $\oplus$ Ekman's Six	**0.59**	**0.52**	0.65	0.80	0.60	0.65	0.77	0.76
$\oplus$ NRC Affect	0.47	0.38	0.69	0.79	0.59	0.67	0.82	0.79
+ Retrofitting $\oplus$ NRC Affect	0.58	0.46	0.72	0.84	**0.62**	**0.71**	**0.85**	**0.82**
+ Counterfitting $\oplus$ NRC Affect	**0.59**	**0.52**	0.65	0.79	0.60	0.65	0.77	0.76
Paragram	0.69	0.54	**0.73**	0.78	0.59	**0.68**	0.80	0.78
$\oplus$ Ekman's Six	0.69	0.55	0.72	0.79	0.59	**0.68**	0.77	**0.79**
+ Retrofitting $\oplus$ Ekman's Six	0.69	0.56	0.72	0.81	0.59	**0.68**	0.80	**0.79**
+ Counterfitting $\oplus$ Ekman's Six	**0.74**	**0.63**	0.69	**0.83**	**0.60**	0.65	**0.85**	0.76
$\oplus$ NRC Affect	0.70	0.55	0.72	0.78	0.59	**0.68**	0.77	**0.79**
+ Retrofitting $\oplus$ NRC Affect	0.69	0.55	0.72	0.80	0.59	**0.68**	0.80	**0.79**
+ Counterfitting $\oplus$ NRC Affect	**0.74**	**0.63**	0.70	0.82	**0.60**	0.65	0.84	0.76

Appendix B Model and Architecture Details

The architecture and hyperparameter details of various models used in the extrinsic evaluation tasks are presented here.

B.1 Sentiment Analysis

We use a Deep Average Network (`https://github.com/miyyer/dan`) with default parameters on binary Stanford Sentiment Treebank (Manning et al., 2014) for this task. Results reported in the paper are accuracies averaged across 5 independent runs.

[13]`http://www.ims.uni-stuttgart.de/forschung/ressourcen/experiment-daten/IMS_emoint_norms.tar.gz`

B.2 Personality Detection

A Convolutional Neural Network(CNN)-based model proposed by Mujumder et al (2017) is used in this paper to evaluate the performance of Aff2Vec embeddigns on the Personality Detection task. We use their Github implementation (`https://github.com/SenticNet/personality-detection`) with *Filter = True, Classifier = MLP, Convolution-Filter = [1,2,3] and Cross-Validation = 10*.

B.3 Emotion Intensity Task (WASSA)

For the WASSA EmoInt-2017, we train a BiLSTM-CNN model with word embeddings as input features. The model is trained separately for each emotion. The network architecture is explained in Table B.1. We use Adam as the optimizer and report Pearson's correlation coefficient(ρ), averaged across 5 independent runs.

Table B.1: BiLSTM-CNN Architecture for WASSA EmoInt-2017

Layer	**Properties**
1D Convolution	$filters = 200, kernel\_size = 3$
Activation	$relu$
1D Max_pooling	$pool\_size = 2$
Dropout	0.3
BiLSTM	$units = 150$
Activation	$relu$
Dropout	0.2
BiLSTM	$units = 80$
Dense	$size = 50$
Activation	$relu$
Dropout	0.3
Dense	$size = 1$

Table B.2: CNN architecture for FFP-Prediction

Layer	**Properties**
2D Convolution	$filters = 5, kernel\_size = 10X5$
Activation	$relu$
2D Max_pooling	$pool\_size = 5X5, sride = 5$
Dense	$size = 50$
Activation	$relu$
Dropout	0.2
Dense	$size = 1$

B.4 Affect Prediction (FFP-Prediction)

We implement a basic CNN-based model for Formality, Frustration and Politeness prediction on Enron-FFP dataset introduced in the paper. The network architecture is shown in Table B.2. We use Rectified Linear Unit (ReLU) as the activation throughout and Stochastic Gradient Descent(SGD) as the optimizer. A mean squared loss is used for regression. We report mean square error averaged across 10 independent runs. Standard deviation for the reported MSE values is of the order of 10^{-3}.