Affect-LM: A Neural Language Model for Customizable Affective Text Generation

Sayan Ghosh[1], Mathieu Chollet[1], Eugene Laksana[1], Louis-Philippe Morency[2] and Stefan Scherer[1]

[1]Institute for Creative Technologies, University of Southern California, CA, USA
[2]Language Technologies Institute, Carnegie Mellon University, PA, USA
[1]{*sghosh,chollet,elaksana,scherer*}*@ict.usc.edu*
[2]*morency@cs.cmu.edu*

Abstract

Human verbal communication includes affective messages which are conveyed through use of emotionally colored words. There has been a lot of research in this direction but the problem of integrating state-of-the-art neural language models with affective information remains an area ripe for exploration. In this paper, we propose an extension to an LSTM (Long Short-Term Memory) language model for generating conversational text, conditioned on affect categories. Our proposed model, *Affect-LM* enables us to customize the degree of emotional content in generated sentences through an additional design parameter. Perception studies conducted using Amazon Mechanical Turk show that *Affect-LM* generates naturally looking emotional sentences without sacrificing grammatical correctness. *Affect-LM* also learns affect-discriminative word representations, and perplexity experiments show that additional affective information in conversational text can improve language model prediction.

1 Introduction

Affect is a term that subsumes emotion and longer term constructs such as mood and personality and refers to the experience of feeling or emotion (Scherer et al., 2010). Picard (1997) provides a detailed discussion of the importance of affect analysis in human communication and interaction. Within this context the analysis of human affect from text is an important topic in natural language understanding, examples of which include sentiment analysis from Twitter (Nakov et al., 2016), affect analysis from poetry (Kao and Jurafsky, 2012) and studies of correlation between function words and social/psychological processes (Pennebaker, 2011). People exchange verbal messages which not only contain syntactic information, but also information conveying their mental and emotional states. Examples include the use of emotionally colored words (such as *furious* and *joy*) and swear words. The automated processing of affect in human verbal communication is of great importance to understanding spoken language systems, particularly for emerging applications such as dialogue systems and conversational agents.

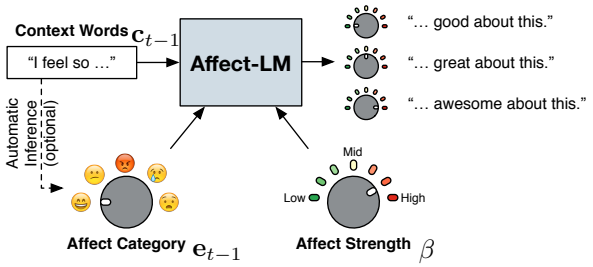

Figure 1: *Affect-LM* is capable of generating emotionally colored conversational text in five specific affect categories ($\mathbf{e}_{t-1}$) with varying affect strengths (β). Three generated example sentences for *happy* affect category are shown in three distinct affect strengths.

Statistical language modeling is an integral component of speech recognition systems, with other applications such as machine translation and information retrieval. There has been a resurgence of research effort in recurrent neural networks for language modeling (Mikolov et al., 2010), which have yielded performances far superior to baseline language models based on n-gram approaches. However, there has not been much effort in building neural language models of text that leverage affective information. Current literature on deep learning for language understanding focuses mainly on representations based on

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 634–642
Vancouver, Canada, July 30 - August 4, 2017.
https://doi.org/10.18653/v1/P17-1059

word semantics (Mikolov et al., 2013), encoder-decoder models for sentence representations (Cho et al., 2015), language modeling integrated with symbolic knowledge (Ahn et al., 2016) and neural caption generation (Vinyals et al., 2015), but to the best of our knowledge there has been no work on augmenting neural language modeling with affective information, or on data-driven approaches to generate emotional text.

Motivated by these advances in neural language modeling and affective analysis of text, in this paper we propose a model for representation and generation of emotional text, which we call the *Affect-LM*. Our model is trained on conversational speech corpora, common in language modeling for speech recognition applications (Bulyko et al., 2007). Figure 1 provides an overview of our *Affect-LM* and its ability to generate emotionally colored conversational text in a number of affect categories with varying affect strengths. While these parameters can be manually tuned to generate conversational text, the affect category can also be automatically inferred from preceding context words. Specifically for model training, the affect category is derived from features generated using keyword spotting from a dictionary of emotional words, such as the LIWC (Linguistic Inquiry and Word Count) tool (Pennebaker et al., 2001). Our primary research questions in this paper are:
Q1: Can *Affect-LM* be used to generate affective sentences for a target emotion with varying degrees of affect strength through a customizable model parameter?
Q2: Are these generated sentences rated as emotionally expressive as well as grammatically correct in an extensive crowd-sourced perception experiment?
Q3: Does the automatic inference of affect category from the context words improve language modeling performance of the proposed *Affect-LM* over the baseline as measured by perplexity?

The remainder of this paper is organized as follows. In Section 2, we discuss prior work in the fields of neural language modeling, and generation of affective conversational text. In Section 3 we describe the baseline LSTM model and our proposed *Affect-LM* model. Section 4 details the experimental setup, and in Section 5, we discuss results for customizable emotional text generation, perception studies for each affect category, and perplexity improvements over the baseline model before concluding the paper in Section 6.

2 Related Work

Language modeling is an integral component of spoken language systems, and traditionally n-gram approaches have been used (Stolcke et al., 2002) with the shortcoming that they are unable to generalize to word sequences which are not in the training set, but are encountered in unseen data. Bengio et al. (2003) proposed neural language models, which address this shortcoming by generalizing through word representations. Mikolov et al. (2010) and Sundermeyer et al. (2012) extend neural language models to a recurrent architecture, where a target word w_t is predicted from a context of all preceding words $w_1, w_2, ..., w_{t-1}$ with an LSTM (Long Short-Term Memory) neural network. There also has been recent effort on building language models conditioned on other modalities or attributes of the data. For example, Vinyals et al. (2015) introduced the neural image caption generator, where representations learnt from an input image by a CNN (Convolutional Neural Network) are fed to an LSTM language model to generate image captions. Kiros et al. (2014) used an LBL model (Log-Bilinear language model) for two applications - image retrieval given sentence queries, and image captioning. Lower perplexity was achieved on text conditioned on images rather than language models trained only on text.

In contrast, previous literature on affective language generation has not focused sufficiently on customizable state-of-the-art neural network techniques to generate emotional text, nor have they quantitatively evaluated their models on multiple emotionally colored corpora. Mahamood and Reiter (2011) use several NLG (natural language generation) strategies for producing affective medical reports for parents of neonatal infants undergoing healthcare. While they study the difference between affective and non-affective reports, their work is limited only to heuristic based systems and do not include conversational text. Mairesse and Walker (2007) developed PERSONAGE, a system for dialogue generation conditioned on extraversion dimensions. They trained regression models on ground truth judge's selections to automatically determine which of the sentences selected by their model exhibit appropriate extroversion attributes. In Keshtkar and Inkpen (2011), the authors use heuristics and rule-based approaches

for emotional sentence generation. Their generation system is not training on large corpora and they use additional syntactic knowledge of parts of speech to create simple affective sentences. In contrast, our proposed approach builds on state-of-the-art approaches for neural language modeling, utilizes no syntactic prior knowledge, and generates expressive emotional text.

3 Model

3.1 LSTM Language Model

Prior to providing a formulation for our proposed model, we briefly describe a LSTM language model. We have chosen this model as a baseline since it has been reported to achieve state-of-the-art perplexities compared to other approaches, such as n-gram models with Kneser-Ney smoothing (Jozefowicz et al., 2016). Unlike an ordinary recurrent neural network, an LSTM network does not suffer from the vanishing gradient problem which is more pronounced for very long sequences (Hochreiter and Schmidhuber, 1997). Formally, by the chain rule of probability, for a sequence of M words $w_1, w_2, ..., w_M$, the joint probability of all words is given by:

$$P(w_1, w_2, ..., w_M) = \prod_{t=1}^{t=M} P(w_t | w_1, w_2, ..., w_{t-1}) \tag{1}$$

If the vocabulary consists of V words, the conditional probability of word w_t as a function of its context $\mathbf{c_{t-1}} = (w_1, w_2, ..., w_{t-1})$ is given by:

$$P(w_t = i | \mathbf{c_{t-1}}) = \frac{\exp(\mathbf{U_i}^T \mathbf{f}(\mathbf{c_{t-1}}) + b_i)}{\sum_{j=1}^{V} \exp(\mathbf{U_j}^T \mathbf{f}(\mathbf{c_{t-1}}) + b_j)} \tag{2}$$

$\mathbf{f}(.)$ is the output of an LSTM network which takes in the context words $w_1, w_2, ..., w_{t-1}$ as inputs through one-hot representations, $\mathbf{U}$ is a matrix of word representations which on visualization we have found to correspond to POS (Part of Speech) information, while $\mathbf{b_i}$ is a bias term capturing the unigram occurrence of word i. Equation 2 expresses the word w_t as a function of its context for a LSTM language model which does not utilize any additional affective information.

3.2 Proposed Model: *Affect-LM*

The proposed model *Affect-LM* has an additional energy term in the word prediction, and can be described by the following equation:

$$P(w_t = i | \mathbf{c_{t-1}}, \mathbf{e_{t-1}}) = \frac{\exp\left(\mathbf{U_i}^T \mathbf{f}(\mathbf{c_{t-1}}) + \beta \mathbf{V_i}^T \mathbf{g}(\mathbf{e_{t-1}}) + b_i\right)}{\sum_{j=1}^{V} \exp(\mathbf{U_j}^T \mathbf{f}(\mathbf{c_{t-1}}) + \beta \mathbf{V_j}^T \mathbf{g}(\mathbf{e_{t-1}}) + b_j)} \tag{3}$$

$\mathbf{e_{t-1}}$ is an input vector which consists of affect category information obtained from the words in the context during training, and $\mathbf{g}(.)$ is the output of a network operating on $\mathbf{e_{t-1}}$.$\mathbf{V_i}$ is an embedding learnt by the model for the i-th word in the vocabulary and is expected to be discriminative of the affective information conveyed by each word. In Figure 4 we present a visualization of these affective representations.

The parameter β defined in Equation 3, which we call the *affect strength* defines the influence of the affect category information (frequency of emotionally colored words) on the overall prediction of the target word w_t given its context. We can consider the formulation as an energy based model (EBM), where the additional energy term captures the degree of correlation between the predicted word and the affective input (Bengio et al., 2003).

3.3 Descriptors for Affect Category Information

Our proposed model learns a generative model of the next word w_t conditioned not only on the previous words $w_1, w_2, ..., w_{t-1}$ but also on the *affect category* $\mathbf{e_{t-1}}$ which is additional information about emotional content. During model training, the affect category is inferred from the context data itself. Thus we define a suitable feature extractor which can utilize an affective lexicon to *infer* emotion in the context. For our experiments, we have utilized the Linguistic Inquiry and Word Count (LIWC) text analysis program for feature extraction through keyword spotting. Introduced by Pennebaker et al. (2001), LIWC is based on a dictionary, where each word is assigned to a predefined LIWC category. The categories are chosen based on their association with social, affective, and cognitive processes. For example, the dictionary word *worry* is assigned to LIWC category *anxiety*. In our work, we have utilized all word categories of LIWC corresponding to affective processes: *positive emotion, angry, sad, anxious*, and *negative emotion*. Thus the descriptor $\mathbf{e_{t-1}}$ has five features with each feature denoting

Corpus Name	Conversations	Words	% Colored Words	Content
Fisher	11700	21167581	3.79	Conversations
DAIC	688	677389	5.13	Conversations
SEMAINE	959	23706	6.55	Conversations
CMU-MOSI	93	26121	6.54	Monologues

Table 1: Summary of corpora used in this paper. CMU-MOSI and SEMAINE are observed to have higher emotional content than Fisher and DAIC corpora.

presence or absence of a specific emotion, which is obtained by binary thresholding of the features extracted from LIWC. For example, the affective representation of the sentence *i will fight in the war* is $\mathbf{e_{t-1}}$ ={"sad":0, "angry":1, "anxiety":0, "negative emotion":1, "positive emotion":0}.

3.4 *Affect-LM* for Emotional Text Generation

Affect-LM can be used to generate sentences conditioned on the input affect category, the affect strength β, and the context words. For our experiments, we have chosen the following affect categories - *positive emotion, anger, sad, anxiety*, and *negative emotion* (which is a superclass of anger, sad and anxiety). As described in Section 3.2, the affect strength β defines the degree of dominance of the affect-dependent energy term on the word prediction in the language model, consequently after model training we can change β to control the degree of how "emotionally colored" a generated utterance is, varying from $\beta = 0$ (neutral; baseline model) to $\beta = \infty$ (the generated sentences only consist of emotionally colored words, with no grammatical structure).

When *Affect-LM* is used for generation, the affect categories could be either (1) inferred from the context using LIWC (this occurs when we provide sentence beginnings which are emotionally colored themselves), or (2) set to an input emotion descriptor $\mathbf{e}$ (this is obtained by setting $\mathbf{e}$ to a binary vector encoding the desired emotion and works even for neutral sentence beginnings). Given an initial starting set of M words $w_1, w_2, ..., w_M$ to complete, affect strength β, and the number of words N to generate each i-th generated word is obtained by sampling from $P(w_i|w_1, w_2, ..., w_{i-1}, \mathbf{e}; \beta)$ for $i \in \{M+1, M+2, ..., M+N\}$.

4 Experimental Setup

In Section 1, we have introduced three primary research questions related to the ability of the proposed *Affect-LM* model to generate emotionally colored conversational text without sacrificing grammatical correctness, and to obtain lower perplexity than a baseline LSTM language model when evaluated on emotionally colored corpora. In this section, we discuss our experimental setup to address these questions, with a description of *Affect-LM*'s architecture and the corpora used for training and evaluating the language models.

4.1 Speech Corpora

The Fisher English Training Speech Corpus is the main corpus used for training the proposed model, in addition to which we have chosen three emotionally colored conversational corpora. A brief description of each corpus is given below, and in Table 1, we report relevant statistics, such as the total number of words, along with the fraction of emotionally colored words (those belonging to the LIWC affective word categories) in each corpus.

Fisher English Training Speech Parts 1 & 2: The Fisher dataset (Cieri et al., 2004) consists of speech from telephonic conversations of 10 minutes each, along with their associated transcripts. Each conversation is between two strangers who are requested to speak on a randomly selected topic from a set. Examples of conversation topics are *Minimum Wage*, *Time Travel* and *Comedy*.

Distress Assessment Interview Corpus (DAIC): The DAIC corpus introduced by Gratch (2014) consists of 70+ hours of dyadic interviews between a human subject and a virtual human, where the virtual human asks questions designed to diagnose symptoms of psychological distress in the subject such as depression or PTSD (Post Traumatic Stress Disorder).

SEMAINE dataset: SEMAINE (McKeown et al., 2012) is a large audiovisual corpus consisting of interactions between subjects and an operator simulating a SAL (Sensitive Artificial Listener). There are a total of 959 conversations which are approximately 5 minutes each, and are transcribed and annotated with affective dimensions.

Multimodal Opinion-level Sentiment Intensity Dataset (CMU-MOSI): (Zadeh et al., 2016) This is a multimodal annotated corpus of opinion

videos where in each video a speaker expresses his opinion on a commercial product. The corpus consist of speech from 93 videos from 89 distinct speakers (41 male and 48 female speakers). This corpus differs from the others since it contains monologues rather than conversations.

While we find that all corpora contain spoken language, they have the following characteristics different from the Fisher corpus: (1) More emotional content as observed in Table 1, since they have been generated through a human subject's spontaneous replies to questions designed to generate an emotional response, or from conversations on emotion-inducing topics (2) Domain mismatch due to recording environment (for example, the DAIC corpus was created in a mental health setting, while the CMU-MOSI corpus consisted of opinion videos uploaded online). (3) Significantly smaller than the Fisher corpus, which is 25 times the size of the other corpora combined. Thus, we perform training in two separate stages - training of the baseline and *Affect-LM* models on the Fisher corpus, and subsequent adaptation and fine-tuning on each of the emotionally colored corpora.

4.2 *Affect-LM* Neural Architecture

For our experiments, we have implemented a baseline LSTM language model in Tensorflow (Abadi et al., 2016), which follows the non-regularized implementation as described in Zaremba et al. (2014) and to which we have added a separate energy term for the affect category in implementing *Affect-LM*. We have used a vocabulary of 10000 words and an LSTM network with 2 hidden layers and 200 neurons per hidden layer. The network is unrolled for 20 time steps, and the size of each minibatch is 20. The affect category $\mathbf{e_{t-1}}$ is processed by a multi-layer perceptron with a single hidden layer of 100 neurons and sigmoid activation function to yield $\mathbf{g(e_{t-1})}$. We have set the output layer size to 200 for both $\mathbf{f(c_{t-1})}$ and $\mathbf{g(e_{t-1})}$. We have kept the network architecture constant throughout for ease of comparison between the baseline and *Affect-LM*.

4.3 Language Modeling Experiments

Affect-LM can also be used as a language model where the next predicted word is estimated from the words in the context, along with an affect category extracted from the context words themselves (instead of being encoded externally as in generation). To evaluate whether additional emotional information could improve the prediction performance, we train the corpora detailed in Section 4.1 in two stages as described below:

(1) **Training and validation of the language models on Fisher dataset**- The Fisher corpus is split in a 75:15:10 ratio corresponding to the training, validation and evaluation subsets respectively, and following the implementation in Zaremba et al. (2014), we train the language models (both the baseline and *Affect-LM*) on the training split for 13 epochs, with a learning rate of 1.0 for the first four epochs, and the rate decreasing by a factor of 2 after every subsequent epoch. The learning rate and neural architecture are the same for all models. We validate the model over the affect strength $\beta \in [1.0, 1.5, 1.75, 2.0, 2.25, 2.5, 3.0]$. The best performing model on the Fisher validation set is chosen and used as a seed for subsequent adaptation on the emotionally colored corpora.

(2) **Fine-tuning the seed model on other corpora**- Each of the three corpora - CMU-MOSI, DAIC and SEMAINE are split in a 75:15:10 ratio to create individual training, validation and evaluation subsets. For both the baseline and *Affect-LM*, the best performing model from Stage 1 (the seed model) is fine-tuned on each of the training corpora, with a learning rate of 0.25 which is constant throughout, and a validation grid of $\beta \in [1.0, 1.5, 1.75, 2.0]$. For each model adapted on a corpus, we compare the perplexities obtained by *Affect-LM* and the baseline model when evaluated on that corpus.

4.4 Sentence Generation Perception Study

We assess *Affect-LM's* ability to generate emotionally colored text of varying degrees without severely deteriorating grammatical correctness, by conducting an extensive perception study on Amazon's Mechanical Turk (MTurk) platform. The MTurk platform has been successfully used in the past for a wide range of perception experiments and has been shown to be an excellent resource to collect human ratings for large studies (Buhrmester et al., 2011). Specifically, we generated more than 200 sentences for four sentence beginnings (namely the three sentence beginnings listed in Table 2 as well as an end of sentence token indicating that the model should generate a new sentence) in five affect categories *happy(positive emotion), angry, sad, anxiety*, and *negative emotion*. The *Affect-LM* model trained

Beginning	Affect Category	Completed sentence
I feel so	Happy	good because i think that it's important to have a relationship with a friend
	Angry	bad that i hate it and i hate that because they they kill themselves and then they fight
	Sad	sad to miss because i i miss the feelings of family members who i lost feelings with
	Anxious	horrible i mean i think when we're going to you know war and alert alert and we're actually gonna die
	Neutral	bad if i didn't know that the decision was going on
I told him to	Happy	be honest and i said well i hope that i 'm going to be a better person
	Angry	see why he was fighting with my son
	Sad	leave the house because i hurt one and i lost his leg and hurt him
	Anxious	be afraid of him and he he just he just didn't care about the death penalty
	Neutral	do this position i think he is he's got a lot of money he has to pay himself a lot of money
Why did you	Happy	have a best friend
	Angry	say it was only a criminal being killed at a war or something
	Sad	miss your feelings
	Anxious	worry about fear factor
	Neutral	believe in divorce

Table 2: Example sentences generated by the model conditioned on different affect categories

on the Fisher corpus was used for sentence generation. Each sentence was evaluated by two human raters that have a minimum approval rating of 98% and are located in the United States. The human raters were instructed that the sentences should be considered to be taken from a conversational rather than a written context: repetitions and pause fillers (e.g., *um, uh*) are common and no punctuation is provided. The human raters evaluated each sentence on a seven-point Likert scale for the five affect categories, overall *affective valence* as well as the sentence's *grammatical correctness* and were paid 0.05USD per sentence. We measured inter-rater agreement using Krippendorffs α and observed considerable agreement between raters across all categories (e.g., for *valence* $\alpha = 0.510$ and *grammatical correctness* $\alpha = 0.505$).

For each *target emotion* (i.e., intended emotion of generated sentences) we conducted an initial MANOVA, with human ratings of affect categories the DVs (dependent variables) and the affect strength parameter β the IV (independent variable). We then conducted follow-up univariate ANOVAs to identify which DV changes significantly with β. In total we conducted 5 MANOVAs and 30 follow-up ANOVAs, which required us to update the significance level to $p<0.001$ following a Bonferroni correction.

5 Results

5.1 Generation of Emotional Text

In Section 3.4 we have described the process of sampling text from the model conditioned on input affective information (research question **Q1**). Table 2 shows three sentences generated by the model for input sentence beginnings *I feel so ...*, *Why did you ...* and *I told him to ...* for each of five affect categories - *happy(positive emotion), angry, sad anxiety*, and *neutral(no emotion)*. They have been selected from a pool of 20 generated sentences for each category and sentence beginning.

5.2 MTurk Perception Experiments

In the following we address research question **Q2** by reporting the main statistical findings of our MTurk study, which are visualized in Figures 2 and 3.

Positive Emotion Sentences. The multivariate result was significant for *positive emotion* generated sentences (Pillai's Trace=.327, $F(4,437)=6.44$, $p<.0001$). Follow up ANOVAs revealed significant results for all DVs except *angry* with $p<.0001$, indicating that both *affective valence* and *happy* DVs were successfully manipulated with β, as seen in Figure 2(a). *Grammatical correctness* was also significantly influenced by the affect strength parameter β and results show that the correctness deteriorates with increasing β (see Figure 3). However, a post-hoc Tukey test revealed that only the highest β value shows a significant drop in *grammatical correctness* at $p<.05$.

Negative Emotion Sentences. The multivariate result was significant for *negative emotion* generated sentences (Pillai's Trace=.130, $F(4,413)=2.30$, $p<.0005$). Follow up ANOVAs revealed significant results for *affective valence* and *happy* DVs with $p<.0005$, indicating that the *affective valence* DV was successfully manipulated with β, as seen in Figure 2(b). Further, as intended there were no significant differences for DVs *angry, sad* and *anxious*, indicating that the *negative emotion* DV refers to a more general affect related concept rather than a specific negative emotion. This finding is in concordance with the intended LIWC category of *negative affect* that forms a parent category above the more

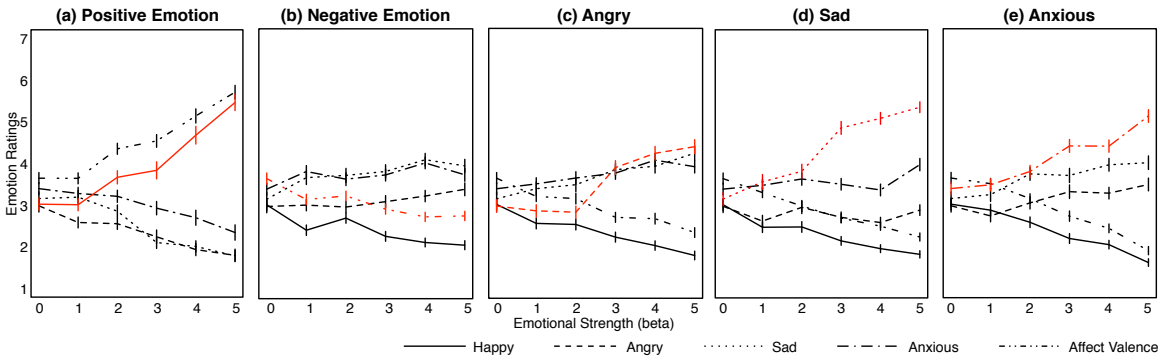

Figure 2: Amazon Mechanical Turk study results for generated sentences in the target affect categories *positive emotion, negative emotion, angry, sad*, and *anxious* (a)-(e). The most relevant human rating curve for each generated emotion is highlighted in red, while less relevant rating curves are visualized in black. Affect categories are coded via different line types and listed in legend below figure.

specific emotions, such as *angry, sad*, and *anxious* (Pennebaker et al., 2001). *Grammatical correctness* was also significantly influenced by the affect strength β and results show that the correctness deteriorates with increasing β (see Figure 3). As for *positive emotion*, a post-hoc Tukey test revealed that only the highest β value shows a significant drop in *grammatical correctness* at p<.05.

Angry Sentences. The multivariate result was significant for *angry* generated sentences (Pillai's Trace=.199, F(4,433)=3.76, p<.0001). Follow up ANOVAs revealed significant results for *affective valence, happy*, and *angry* DVs with p<.0001, indicating that both *affective valence* and *angry* DVs were successfully manipulated with β, as seen in Figure 2(c). *Grammatical correctness* was not significantly influenced by the affect strength parameter β, which indicates that angry sentences are highly stable across a wide range of β (see Figure 3). However, it seems that human raters could not successfully distinguish between *angry, sad*, and *anxious* affect categories, indicating that the generated sentences likely follow a general *negative affect* dimension.

Sad Sentences. The multivariate result was significant for *sad* generated sentences (Pillai's Trace=.377, F(4,425)=7.33, p<.0001). Follow up ANOVAs revealed significant results only for the *sad* DV with p<.0001, indicating that while the *sad* DV can be successfully manipulated with β, as seen in Figure 2(d). The *grammatical correctness* deteriorates significantly with β. Specifically, a post-hoc Tukey test revealed that only the two highest β values show a significant drop in *grammatical correctness* at p<.05 (see Figure 3).

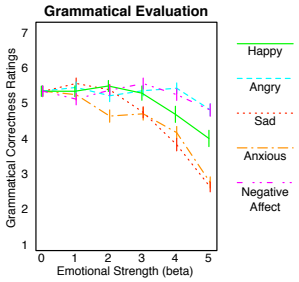

Figure 3: Mechanical Turk study results for *grammatical correctness* for all generated target emotions. Perceived *grammatical correctness* for each affect categories are color-coded.

A post-hoc Tukey test for *sad* reveals that $\beta = 3$ is optimal for this DV, since it leads to a significant jump in the perceived sadness scores at p<.005 for $\beta \in \{0, 1, 2\}$.

Anxious Sentences. The multivariate result was significant for *anxious* generated sentences (Pillai's Trace=.289, F(4,421)=6.44, p<.0001). Follow up ANOVAs revealed significant results for *affective valence, happy* and *anxious* DVs with p<.0001, indicating that both *affective valence* and *anxiety* DVs were successfully manipulated with β, as seen in Figure 2(e). *Grammatical correctness* was also significantly influenced by the affect strength parameter β and results show that the correctness deteriorates with increasing β. Similarly for *sad*, a post-hoc Tukey test revealed that only the two highest β values show a significant drop in *grammatical correctness* at p<.05 (see Figure 3). Again, a post-hoc Tukey test for *anxious* reveals that $\beta = 3$ is optimal for this DV, since it leads to a significant jump in the perceived

	Training (Fisher)		Adaptation	
Perplexity	**Baseline**	**Affect-LM**	**Baseline**	**Affect-LM**
Fisher	37.97	37.89	-	-
DAIC	65.02	64.95	55.86	55.55
SEMAINE	88.18	86.12	57.58	57.26
CMU-MOSI	104.74	101.19	66.72	64.99
Average	73.98	72.54	60.05	59.26

Table 3: Evaluation perplexity scores obtained by the baseline and *Affect-LM* models when trained on Fisher and subsequently adapted on DAIC, SEMAINE and CMU-MOSI corpora

anxiety scores at p<.005 for $\beta \in \{0, 1, 2\}$.

5.3 Language Modeling Results

In Table 3, we address research question **Q3** by presenting the perplexity scores obtained by the baseline model and *Affect-LM*, when trained on the Fisher corpus and subsequently adapted on three emotional corpora (each adapted model is individually trained on CMU-MOSI, DAIC and SEMAINE). The models trained on Fisher are evaluated on all corpora while each adapted model is evaluated only on it's respective corpus. For all corpora, we find that *Affect-LM* achieves lower perplexity on average than the baseline model, implying that affect category information obtained from the context words improves language model prediction. The average perplexity improvement is 1.44 (relative improvement 1.94%) for the model trained on Fisher, while it is 0.79 (1.31%) for the adapted models. We note that larger improvements in perplexity are observed for corpora with higher content of emotional words. This is supported by the results in Table 3, where *Affect-LM* obtains a larger reduction in perplexity for the CMU-MOSI and SEMAINE corpora, which respectively consist of 2.76% and 2.75% more emotional words than the Fisher corpus.

5.4 Word Representations

In Equation 3, *Affect-LM* learns a weight matrix $\mathbf{V}$ which captures the correlation between the predicted word w_t, and the affect category $\mathbf{e_{t-1}}$. Thus, each row of the matrix $\mathbf{V_i}$ is an emotionally meaningful embedding of the i-th word in the vocabulary. In Figure 4, we present a t-SNE visualization of these embeddings, where each data point is a separate word, and words which appear in the LIWC dictionary are colored based on which affect category they belong to (we have labeled only words in categories *positive emotion*, *negative emotion*, *anger*, *sad* and *anxiety* since

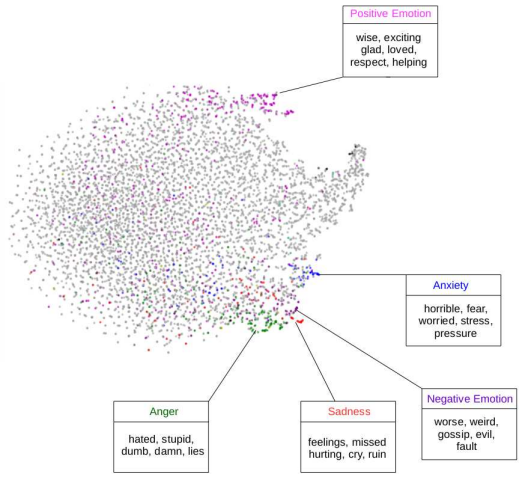

Figure 4: Embeddings learnt by *Affect-LM*

these categories contain the most frequent words). Words colored grey are those not in the LIWC dictionary. In Figure 4, we observe that the embeddings contain affective information, where the positive emotion is highly separated from the negative emotions (*sad*, *angry*, *anxiety*) which are clustered together.

6 Conclusions and Future Work

In this paper, we have introduced a novel language model *Affect-LM* for generating affective conversational text conditioned on context words, an affective category and an affective strength parameter. MTurk perception studies show that the model can generate expressive text at varying degrees of emotional strength without affecting grammatical correctness. We also evaluate *Affect-LM* as a language model and show that it achieves lower perplexity than a baseline LSTM model when the affect category is obtained from the words in the context. For future work, we wish to extend this model by investigating language generation conditioned on other modalities such as facial images and speech, and to applications such as dialogue generation for virtual agents.

Acknowledgments

This material is based upon work supported by the U.S. Army Research Laboratory under contract number W911NF-14-D-0005. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Government, and no official endorsement should be inferred. Sayan Ghosh also acknowledges the Viterbi Graduate School Fellowship for funding his graduate studies.

References

Sungjin Ahn, Heeyoul Choi, Tanel Pärnamaa, and Yoshua Bengio. 2016. A neural knowledge language model. *arXiv preprint arXiv:1608.00318* .

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research* 3(Feb):1137–1155.

Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. 2011. Amazon's mechanical turk a new source of inexpensive, yet high-quality, data? *Perspectives on psychological science* 6(1):3–5.

Ivan Bulyko, Mari Ostendorf, Manhung Siu, Tim Ng, Andreas Stolcke, and Özgür Çetin. 2007. Web resources for language modeling in conversational speech recognition. *ACM Transactions on Speech and Language Processing (TSLP)* 5(1):1.

Kyunghyun Cho, Aaron Courville, and Yoshua Bengio. 2015. Describing multimedia content using attention-based encoder-decoder networks. *IEEE Transactions on Multimedia* 17(11):1875–1886.

Christopher Cieri, David Miller, and Kevin Walker. 2004. The fisher corpus: a resource for the next generations of speech-to-text. In *LREC*. volume 4, pages 69–71.

Martín Abadi et al. 2016. Tensorflow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI). Savannah, Georgia, USA*.

Jonathan et al. Gratch. 2014. The distress analysis interview corpus of human and computer interviews. In *LREC*. Citeseer, pages 3123–3128.

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

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410* .

Justine Kao and Dan Jurafsky. 2012. A computational analysis of style, affect, and imagery in contemporary poetry.

Fazel Keshtkar and Diana Inkpen. 2011. A pattern-based model for generating text to express emotion. In *Affective Computing and Intelligent Interaction*, Springer, pages 11–21.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014. Multimodal neural language models.

Saad Mahamood and Ehud Reiter. 2011. Generating affective natural language for parents of neonatal infants. In *Proceedings of the 13th European Workshop on Natural Language Generation*. Association for Computational Linguistics, pages 12–21.

François Mairesse and Marilyn Walker. 2007. Personage: Personality generation for dialogue.

Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 2012. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing* 3(1):5–17.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Interspeech*. volume 2, page 3.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. pages 3111–3119.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. Semeval-2016 task 4: Sentiment analysis in twitter. *Proceedings of SemEval* pages 1–18.

James W Pennebaker. 2011. The secret life of pronouns. *New Scientist* 211(2828):42–45.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates* 71(2001):2001.

Rosalind Picard. 1997. *Affective computing*, volume 252. MIT press Cambridge.

Klaus R Scherer, Tanja Bänziger, and Etienne Roesch. 2010. *A Blueprint for Affective Computing: A sourcebook and manual*. Oxford University Press.

Andreas Stolcke et al. 2002. Srilm-an extensible language modeling toolkit. In *Interspeech*. volume 2002, page 2002.

Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. 2012. Lstm neural networks for language modeling. In *Interspeech*. pages 194–197.

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

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems* 31(6):82–88.

Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. 2014. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329* .