INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

EmoStory: Emotion Prediction and Mapping in Narrative Stories

Seng-Wei Too [a], John See [b], Albert Quek [c], Hui-Ngo Goh [c,*]

[a] Auronex Sdn Bhd, Bandar Bukit Jalil, 57000, Kuala Lumpur, Malaysia
[b] School of Mathematical and Computer Sciences, Heriot-Watt University Malaysia, Putrajaya, Malaysia
[c] Faculty of Computing and Informatics, Multimedia University, Cyberjaya, 63100, Malaysia.
Corresponding author: *hngoh@mmu.edu.my

***Abstract*—A well-designed story is built upon a sequence of plots and events. Each event has its purpose in piquing the audience's interest in the plot; thus, understanding the flow of emotions within the story is vital to its success. A story is usually built up through dramatic changes in emotion and mood to create resonance with the audience. The lack of research in this understudied field warrants exploring several aspects of the emotional analysis of stories. In this paper, we propose an encoder-decoder framework to perform sentence-level emotion recognition of narrative stories on both dimensional and categorical aspects, achieving MAE=0.0846 and 54% accuracy (8-class), respectively, on the EmoTales dataset and a reasonably good level of generalization to an untrained dataset. The first use of attention and multi-head attention mechanisms for emotion representation mapping (ERM) yields state-of-the-art performance in certain settings. We further present the preliminary idea of EmoStory, a concept that seamlessly predicts both dimensional and categorical space in an efficient manner, made possible with ERM. This methodology is useful in only one of the two aspects is available. In the future, these techniques could be extended to model the personality or emotional state of characters in stories, which could benefit the affective assessment of experiences and the creation of emotive avatars and virtual worlds.**

***Keywords*— Deep learning; affective computing; natural language processing.**

Manuscript received 30 Dec. 2022; revised 29 Jun. 2021; accepted 2 Sep. 2023. Date of publication 30 Nov. 2023.

I. INTRODUCTION

In storytelling, a well-written story is key to creating an emotional resonance with the audience without the support of external mediums such as video and audio. A well-designed story is built upon a sequence of plots and events. Normally, each event has its purpose in piquing the audience's interest in the plot; thus, understanding the flow of emotions within the story is particularly vital to its success. For instance, the transition from one plot to another, such as from the exposition to the climax, should be well structured to immerse the audience into the emotional flow of these events. A sample snippet from the Cinderella fairy tale story and its associated emotional states is illustrated in Figure 1. It can be observed that just within a few lines, the story has undergone a dramatic change in emotion and mood.

Generally, story-based emotion recognition aims to interpret the human effect from a narrative story [1], and it should be done based on the homologous structure of emotions and narratives [2]. Emotions can be quantified in two different spaces: (i) categorical space and (ii) dimensional space. Motion in categorical space is simply a pre-defined set of emotions that we use in daily life, such as “happy” and “sad” while emotion in dimensional space, such as the VAD, i.e., valence (V), arousal (A) and dominance (D), describes emotion in three axes. According to Beuchel and Hahn [3], the continuous values of dimensional space provide a better description than the categorical space, which is limited by the available classes. VAD is not limited to emotion recognition but is also observable in other domains. Ghosh et al. [4] utilized the VAD model in their study on suicide notes to detect both emotion class and intensity from sentences, resulting in improved overall performance. The VAD model can help determine a person's exact emotion(s) and its intensity, which is an important clue about suicidal tendencies.

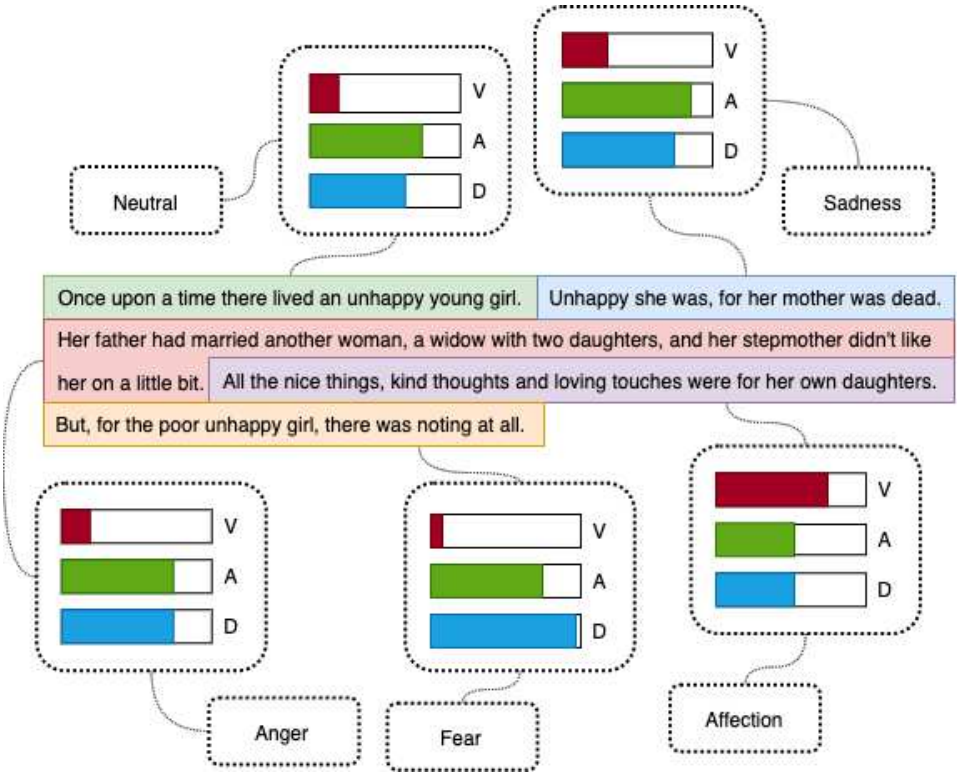

Fig. 1 The emotion flow from an excerpt from 'Cinderella'. A single passage can contain a dramatic change in emotions

Liu et al. [5] proposed the use of the modified Plutchik's wheel of emotions (consisting of 8 primary emotions) to predict the emotion conveyed in a fairy tale passage [6]. Although the dataset is quite large and multi-class labels were used, their work was performed at the passage level (averaging 6.24 sentences per passage), which may not precisely capture the complex emotional flow of a storyline. This causes an inevitable loss of emotional information within a full passage. Emotion Representation Mapping (ERM) is a task to map between dimensional and categorical space.

1) Emotion models: The categorical model is one of the most natural human-like approaches to interpreting emotions, and it is based on a pre-defined set of emotional categories. However, this model limits the level of precision, and it is difficult to present a complex emotion as the number of choices available restricts it. Five basic emotions are widely adopted: happiness, sadness, fear, anger, and disgust [7], which are also 6 [8] or 9 [9] in various works. Another work proposed OntoEmotions [10], a hierarchical ontology of emotion consisting of a pre-selected 119 emotions further clustered into nine basic emotions. Work proposed by [11] recognizes emotion using facial expressions and voice. Dimensional models consider affective states to be best described by a small number of independent emotional concepts. The VAD model, being one of the most popularly used in research [3], consists of three axes, where each ax represents a different concept: Valence (positiveness or negativeness of an emotion), Arousal (calm-excited scale) and Dominance (indicative of whether the subject feels in control of the situation or not) [12].

2) Emotion Representation Mapping (ERM): As both emotional models cover different attributes, ERM is a task that converts an emotional rating from one representative to another one, e.g., mapping the emotion categories to VAD. Buechel and Hahn [13] proposed a multi-task MLP approach to ERM that surpasses the previous state-of-the-art, achieving close to human-level performance. Their neural network entails two fully connected layers (128 units with ReLU activation) with two dropouts (0.2) following each layer. Their paper claims that their methodology is independent of the language barrier (workable under cross-lingual settings). However, the architecture seemed too simple to understand the complex emotional mapping [14].

3) Emotion Recognition in Text: The task of identifying emotion from text has been an active research area for over a decade, with numerous works focusing on news headlines [8], blog posts [9], and tweets [15]. In recent years, researchers working in the NLP domain have explored deep learning approaches to great success. Approaches such as ELMO [16] and BERT [17] learned deep bidirectional word representations that are robust across a broad range of NLP problems; the latter is particularly advantageous in its use of unlabeled data with the popular Transformer architecture [18]. The transformer consists of only self-attention mechanisms, bypassing the need for an RNN. Recent research is increasingly gravitating towards attention-based methods—several recent approaches designed attentional mechanisms for emotion regression [19] and ranking [20]. Liu et al. [21] employed the method of fine-tuning BERT for emotion prediction at the passage level (multiple sentences). While they justify that some sentences may not carry any emotion, this also afflicts a loss of valuable emotional context propagated through the passage. Another recent work by Wu et al. [22] proposed a Memory Fusion Network (attention across multiple modalities) and a Transformer network to model emotions from narrative video. We take a cue from this work that processing at the sentence level is a promising

direction. Cortiz [23] compared the performance of different pre-trained transformer models, including BERT, DistilBERT, RoBERTa, XLNet, and ELECTRA, in detecting emotions in texts and found that RoBERTa presented the best metrics for accuracy and macro-f.

4) Emotion Representation Mapping (ERM): ERM is a task to perform the mapping between dimensional space and categorical space. To facilitate scenarios where labels are only in dimensional or categorical space, ERM is a linkage step for translation between spaces. Work by Buechel and Hahn [13] presented a multi-task multilayer perceptron (MLP) to perform cross-lingual and monolingual ERM from the basic five emotions to VAD and vice versa. However, their simplistic model may not impose the emotional mapping well at the sentence-level emotion. The key problems of this work are:

- To improve existing ERM to generalize in cross-lingual learning settings, as most datasets are cross-lingual instead of monolingual.
- To propose a framework that retains direct recurrency between words in a sentence level.
- To experiment with using ERM in the emotion recognition model in dimensional and categorical spaces.

The key contributions of this paper are as follows:

- We introduce the first use of attention and multi-head attention mechanisms for emotion representation mapping (ERM), yielding state-of-the-art performance.
- We propose an encoder-decoder framework to perform sentence-level emotion recognition of narrative stories, which is feasible in both dimensional and categorical space, achieving MAE=0.0846 and 54% accuracy (8-class), respectively, on EmoTales dataset.
- We trialed the preliminary idea of EmoStory, a concept that seamlessly predicts both dimensional and categorical space in an efficient manner with the aid of ERM. This methodology is useful in only one of the two aspects is available.

II. Materials and Method

Here, we describe the main dataset used for the problems and solutions proposed in this work.

A. Datasets

EmoTales [10] is designed for affective computing in the narrative story domain. The corpus comprises 1,389 English sentences from 18 distinct fairy tales with 3,621 unique English words. The average word count per sentence is 15, and the maximum word count per sentence is 100. The label contains both categorical and dimensional schema. One of the main challenges is dealing with the data shortage in the dataset. EmoTales is annotated by 36 people for the categorical model (119 fine categories were annotated, which were further narrowed down to 9 categories using OntoEmotion [24], while 26 people annotated the dimensional model.

Interestingly, fairy tales are selected in this dataset as they are generally intended to help children better understand and experience feelings on their way to maturity. EmoTales is annotated according to the Self-Assessment Manikin (SAM) standard [25], where the range of the dimensional model is mapped into an integer between 1 and 9. SAM ensures that it is not restricted to any one culture and language and is appropriate for use in different scenarios.

120 Stories [26] dataset is a collection of 4,000 short stories for sentiment analysis, which was crawled from a website named American Literature. We utilized only the stories that contained 128 words or less in a passage (~120). Notably, this dataset was collected at the passage level instead of the sentence level with the emotional content calculated using the valence, arousal, and dominance norms according to SAM and further normalized to the range between 0 to 1. The authors created this dataset as part of a study on the impact of narrative literature on mental health.

B. Problem Setting

There are two tasks involved in our work on processing emotions in text.

1) Emotion Recognition: To recognize emotions, assume a sentence, X is sampled from a story/passage, an optimal function $f(X)$ is needed to interpret the emotion context of X in categorical space Yc, or dimensional space, Ye, where the emotion category $Y_c \in C = \{$ *affection, anger, bravery, fear, happiness, sadness, surprise, neutral*$\}$, and $Y_e = [0, 1]$ after normalizing values from [0, 9] (based on SAM) with $e = (v, a, d)$ a triplet representing the valence v, arousal a and dominance d values. Succinctly:

$$Y_c = f_c(X) \tag{1}$$

$$Y_e = Y_{v,a,d} = f_e(X) \tag{2}$$

2) Emotion Representation Mapping. In this task, we aim to find a function $g\ (v, a, d)$ that best maps the dimensional triplet to a single Y_c category and vice versa. The problem is summarized as follows:

$$Y_c = g(e) = g((v, a, d)) \equiv g : e \rightarrow c \tag{3}$$

$$Y_e = \hat{g}\ (c) \equiv \hat{g}: c \ \rightarrow e \tag{4}$$

where g and ĝ are two separate functions that map to opposing directions.

C. Emotion Representation Mapping with Multi-head Self-Attention

We first describe our proposed method for ERM. In this work, we leverage the concept of self-attention to stimulate the mapping from the dimensional schema (i.e., VAD) to the categorical scheme (i.e., one of 8 emotions). The use of attention plays an important role in capturing the nature of human emotion by deciding which dimension should be paid more attention, as with the works of [19], and [22].

By adopting the concept of multi-head attention [18], the architecture learns to understand the pattern of emotion under a different scenario. In other words, each head is responsible for each scenario, which can help to understand complex emotions in cross-lingual settings or from a long sentence. The mapping model $M(\cdot)$, takes an input e with three dimensions (v, a, d) and propagates it through h number of self-attention heads and fully-conected layer. At the output layer, a softmax classifier $S_{erm}(\cdot)$ is applied to predict the probability P_c of each emotion c, and the final mapped category $\hat{Y}_c$ is obtained by:

$$\hat{Y}_c = S_{erm}(e) = {}_{Yc \in C} P_c \quad (5)$$

where

$$Pc = \frac{\exp(W^T(e_c))}{\sum_j^c \exp(W^T(e_j))} \quad (6)$$

and W are parameters learned from model M. Fig. 2 shows the general flow of the Multi-head Attention Network employed for ERM.

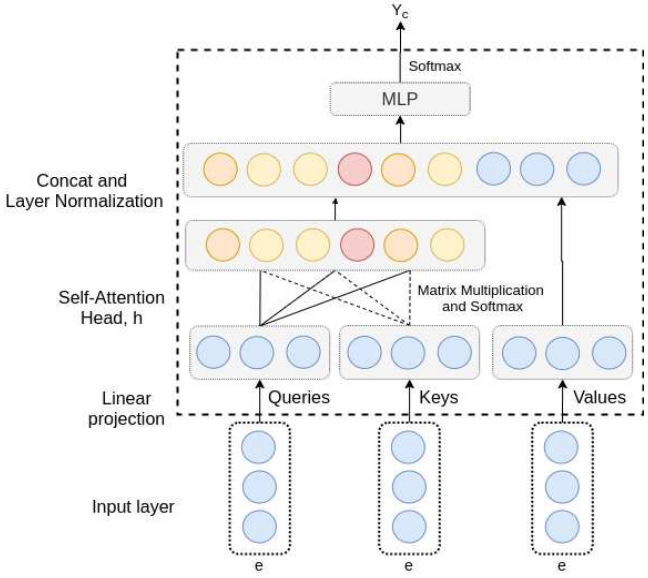

Fig. 2 Multi-head Attention for Emotion Representation Mapping (ERM)

D. Emotion Recognition with Encoder-Decoder Framework

To accomplish emotion recognition, we propose to employ an encoder-decoder framework with the BERT [17] as the potential choice of the encoder (Ɛ) owing to its promising performance as an unsupervised feature extractor. In our work, we adapt the uncased BERT-Large model with L = 12 transformer blocks, a hidden size of H = 768, and A = 12 attention heads. This encoder takes an input text sequence X of no more than 128 words, where each word goes through an embedding before mapping to a hidden state. In each transformer layer, the network learns the relationship between each word in the sentence bi-directionally (as self-attention allows seeing into future information). Each layer in the transformer is responsible for capturing different features. The first few layers of the transformer block capture low-level features such as language vocabulary, while the last few layers capture high-level features such as semantic meaning and emotional context.

Unlike the encoder side, we opt for methods on the decoder side that can retain direct recurrency between words in a sentence. For this, we exploit the Bi-directional LSTM as the choice of the decoder (D) to learn the recurrency between each hidden state by taking two inputs to the hidden layer: (1) the output of the encoder, $\mathcal{E}(X)_t$ at time-step t, and (2) hidden state of the previous time-step, h_{t-1}. We set our decoder side with s = 256 hidden states (a total of 512 hidden states for both directions) to ensure consistency in the output condition for benchmarking various methods. The final hidden state produced by LSTM, h_s, is passed through a fully connected layer to an output layer. The output layer is a softmax classifier $S(\cdot)$ if the desired output is Y_c or a linear regressor $R(\cdot)$ if the desired output is Y_e.

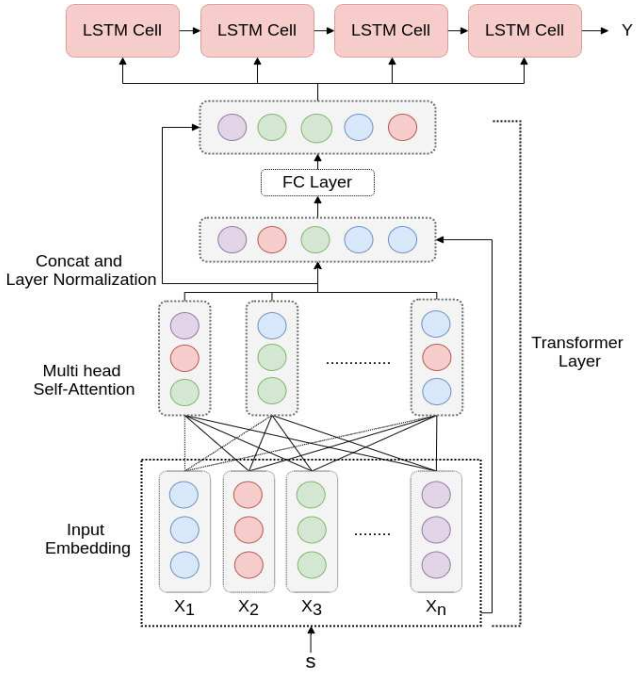

Fig. 3 Encoder-Decoder Framework for emotion recognition

$$Y_c = S(D_c(h_{s-1}, \mathcal{E}(X)_s)) \quad (7)$$

$$Y_e = R(D_e(h_{s-1}, \mathcal{E}(X)_s)) \quad (8)$$

E. Emotion Recognition

With the initial motivation to trial the feasibility of ERM in the emotion recognition task, we propose EmoStory as a framework that aims to seamlessly combine ERM with the encoder-decoder framework to enable the predicted dimensional outputs to be translated directly to its corresponding categorical space and vice versa. This method also promotes efficiency by way of avoiding retraining another large BERT model to predict for the other emotion space. EmoStory combines the concept of ERM with the previously mentioned encoder-decoder framework, f_{vad}. More specifically, EmoStory first takes an input sentence, X, and uses $f_e(X)$ to predict the $Y_{v,a,d}$ based on Eq. 8. Then, it proceeds to propagate the aforementioned multi-head self-attention ERM to obtain the output category, Y_c:

$$Yc = g(f_e(X)) \quad (9)$$

Here, we show details of our experimental setups and a brief description of evaluated methods, starting with the ERM and then emotion recognition.

F. ERM Efficacy

To measure the efficacy of performing ERM, we employ a combination of several metrics as deemed suitable to the datasets used. Pearson's correlation is generally used in ERMDB for both directions of mapping since both the categorical (Basic Emotion 5 (BE5)) and dimensional (VAD) scores are available via Buechel and Hahn [7]. The authors noted that the Pearson correlation is the most consistent and comparable to human performance. We use a similar ERMDB experiment setting for fair benchmarking previously done Buechel and Hahn's work [13]. The F1-score is used only for

EmoTales dataset as the dataset contains emotion categories annotated in discrete or one hot encoded manner.

MLP: MLP with two-hidden layer FFNNs (both with 128 units) and ReLU activation, followed by 2 dropouts for each layer. Their methodology works under cross-lingual settings as well. Their method obtained an almost human-performance level on categorical-to-dimensional and dimensional-to-categorical mapping. (ii) MLP + Multi-head Self Attention (Ours): Before applying the vanilla MLP, a multi-head self-attention module is added to learn the relationship among the input. A linear projection of 256 units for Queries, Keys, Values, and dot product attention is used here. Different numbers of attention heads, H = {1, 2, 4, 8} are tested.

G. Experimentations of Emotion Recognition

We have experimented with different encoder and decoder models combinations for this task. We use the following comparison methods for the encoder: (i) MLP: A single hidden layer of 128 units followed by dropout (0.2). (ii) Attention [27]: Linearly projected 128- dimensional keys, values, and queries. Dot-product self-attention is then applied, followed by a layer normalization process. (iii) Transformer [18]: 4 stacked transformer layers, with each transformer layer consisting of 4 self-attention (scaled dot product attention) modules, which are then concatenated. (iv) BERT [17]: A pre-trained bidirectional language model trained in an unsupervised manner on a large text corpus to learn a representation that can be used to fine-tune other NLP problems. We adapted the uncased BERT-Large model with L = 12 transformer blocks, a hidden size of H = 768, and A = 12 attention heads. (v) RoBERTa [28]: A variant of BERT which modifies a part of the pre-training technique and hyperparameter choices in BERT. We adopted the pretrained version, that has 12 layers, 768 hidden units, and 12 attention heads (similar to BERT setting). (vi) ALBERT [29]: A lightweight version of BERT which is optimized by factorization of the embedding parameters. We adopted the version that contains 12 stacked transformer layers with the embedding size of 128, 768 hidden states, and 12 self-attention heads (similar to BERT setting). (vii) ELMo [16]: A deep contextualized word representation built based on a bi-directional language model. We used the pre-trained ELMo model (v2) available on Tensorhub. We feed the word embeddings of ELMo as input to the decoder.

TABLE I
RESULT OF EMOTION REPRESENTATION MAPPING (ERM) COMPARING THE BASELINE MLP AGAINST VARIANTS OF THE PROPOSED MODEL. THERE ARE 5 EMOTION CLASSES IN ERMDB AND 8 EMOTION CLASSES IN EMOTALES

Model	EmoTales		ERMDB			
	e → c	c → e	e → c		c → e	
	F1-Score	MAE	ρmono	ρcross	ρmono	ρcross
MLP [6]	0.37	0.1003	0.8385	0.8035	0.8402	0.8046
MLP + 1H-SelfAtt	0.59	0.0856	0.8307	0.8004	0.8312	0.8108
MLP + 2H-SelfAtt	0.53	0.0870	0.8228	0.8152	0.8150	0.7907
MLP + 4H-SelfAtt	0.59	0.0845	0.8223	0.8110	0.8164	0.8035
MLP + 8H-SelfAtt	0.51	0.0847	0.8196	0.8127	0.8135	0.7993

For decoder, we experimented with three types of decoder models: (i) FC: A basic single fully connected layer of 128 units followed by dropout (0.2). (ii) LSTM: One-layer uni-directional RNN (256 units) with LSTM cells. (iii) BiLSTM: Like LSTM, except that this is bi-directional, thus allowing the model to learn representation and features from both directions.

H. Implementation Details

All encoder-decoder models are trained using SGD (learning rate of 10^{-3}) while the ERM model is optimized with Adam (learning rate of 5×10^{-4}) EmoTales dataset is split using 60:20:20 ratio for the train-validation-test partitions. To increase the amount of training data, we employed the Easy Data Augmentation (EDA) technique [30] to perform text augmentation followed by Principal Component Analysis (PCA)We manually balance (by over-sampling via augmentation and sub-sampling) the sample distribution for each class, ensuring around 700 samples per class for training. Overall, we also omitted Disgust as its number of samples is too small; hence we have 8 emotion categories: Affection, Anger, Bravery, Fear, Happiness, Neutral, Sadness, and Surprise.

III. RESULTS AND DISCUSSION

We first present comprehensive experiment results for the ERM and emotion recognition tasks, followed by more analysis and discussion.

A. Emotion Representation Mapping

In this work, we first evaluate and compare methods for mapping the dimensional VAD values to the eight emotion categories on the EmoTales and the original lexical datasets used in Beuchel & Hahn [13](which from here on denoted collectively as 'ERMDB') All models are trained with the test inference performed on the same dataset. For benchmarking consistency, we report the Pearson's correlation between the predicted value and ground-truth label/values ρ for ERMDB, while EmoTales retained the use of F1-score and MAE for classification of categories and regression of VAD values, respectively. Specifically, training on ERMDB was conducted using SGD optimizer with a learning rate of $7\times10^{-}4$, while training on EmoTales follows the base setting presented in Section II (H). Table I presents the experimental results of the ERM task on the EmoTales and ERMDB datasets. Further comparisons against the baseline of Beuchel and Hahn [13] are shown in Tables II and III. The baseline experiments are reproduced and reported in both tables following the authors' original configuration (of MLP) in both monolingual and cross-lingual settings. Compared to a

monolingual setting, the proposed framework is generally more robust, especially in the cross-lingual setting. The complexity of sentence-level emotion in EmoTales warrants a method that carries relatively more information; the attentional mechanism showed that it can learn the context and dependencies between the three dimensions of emotion. The proposed multi-head self-attention method helps to generalize to different scenarios in the case of cross-lingual learning.

Besides, our ablation studies in Tables II and III showed that our method's ideal number of attention heads remains inconclusive; performances vary across different scenarios. Generally, the '4H-Attention' variant (literally, 4 attention heads) worked better on both the EmoTales dataset (Table I) and the cross-lingual ERMDB dataset (Table III). Although the baseline method was surprisingly strong in the monolingual ERMDB, the 1H and 2H variants of the proposed method are quite competitive across the board, fairing marginally better than the rest in most scenarios.

TABLE II
ERM BASELINE COMPARISON IN MONOLINGUAL SETTING

Model	V	A	D	Joy	Ang	Sad	Fear	Dis	AvgVAD	AvgBE5	Overall
Baseline-Reported	0.956	0.7540	0.8100	0.932	0.855	0.816	0.838	0.752	0.8400	0.8386	1.6786
***f* Baseline-Reproduced**	**0.9558**	**0.7545**	**0.8104**	**0.9323**	**0.8555**	**0.8155**	**0.8377**	**0.7516**	**0.8402**	**0.8385**	**1.6787**
MLP + 1H-SelfAtt	0.9539	0.7395	0.8002	0.9109	0.8507	0.8116	0.8346	0.7459	0.8312	0.8307	1.6619
MLP + 2H-SelfAtt	0.9487	0.7218	0.7745	0.9204	0.842	0.7985	0.8241	0.7291	0.8150	0.8228	1.6378
MLP + 4H-SelfAtt	0.9486	0.7205	0.7802	0.9199	0.8423	0.7988	0.8247	0.7261	0.8164	0.8223	1.6388
MLP + 8H-SelfAtt	0.9487	0.7143	0.7775	0.9201	0.8403	0.7944	0.8216	0.7216	0.8135	0.8196	1.6331

B. Emotion Recognition

The proposed encoder-decoder framework is designed to interpret emotional context in dimensional and categorical spaces. As both tasks are trained independently and evaluated using a different metric (categorical with F1-score, dimensional with MAE), Table V presents a comprehensive set of experiments comparing different encoder and decoder models, as outlined in Section II(G). It turns out that BERT as the encoder and LSTM as the decoder achieved the best result for categorical emotion prediction with an F1-score of 0.5467. However, EmoStory, which incorporated the concept of ERM (mapping VAD to categories) did not perform as well as expected. We surmise that the conversion from hidden state to VAD resulted in a loss of information and thus dented the performance of the prediction.

TABLE III
ERM BASELINE COMPARISON IN CROSS-LINGUAL SETTING

Model	V	A	Joy	Ang	Sad	Fear	Dis	VAD	BE5	Overall
Baseline-Reported	0.9510	0.6740	0.9260	0.829	0.7930	0.7860	0.7310	0.8125	0.8130	1.6255
f Baseline-Reproduced	**0.9506**	0.6586	0.9019	0.8264	0.7826	0.7874	0.7193	0.8046	0.8035	1.6081
MLP + 1H-SelfAtt	0.9485	**0.6731**	0.9007	0.8258	0.7743	0.7932	0.7080	**0.810**8	0.8004	1.6112
MLP + 2H-SelfAtt	0.9433	0.6382	**0.9160**	0.8373	**0.7920**	**0.8042**	**0.7265**	0.7908	**0.8152**	1.6060
MLP + 4H-SelfAtt	0.9426	0.6645	**0.9160**	0.8365	0.7842	0.7992	0.7192	0.8036	0.8110	**1.6146**
MLP + 8H-SelfAtt	0.9407	0.6580	0.9130	**0.8376**	0.7858	0.8033	0.7240	0.7994	0.8127	1.6121

For VAD emotion prediction task, BERT with a Bidirectional LSTM decoder achieved the best result with MAE= 0.0846 and MSE=0.0122. We also inferred the trained models from EmoTales on a passage-level 120 Stories dataset to examine its capability at generalizing to other corpora. The best-performing method (BERT or RoBERTa + FC) achieved MAE of around 0.1, putting it at ∼10% absolute error off the ground truth (scaled to between 0 and 1). This is evidential that the learned knowledge can be directly used in a slightly different data structure (passage-level), without further retraining necessary. Table IV shows a few sample sentences from the EmoTales test set (casing stripped) together with their corresponding predicted emotion category and ground-truth label.

TABLE IV
SAMPLE SENTENCES FROM THE EMOTALES TEST SET WITH THEIR TRUE AND PREDICTED LABELS

#	Test Sample	True	Predicted
1	she gave him the mirror in his hand, and he saw there in the likeness of the most beautiful maiden on earth	surprise	happiness
2	if that is the ladder by which one mounts, i too will try my fortune,	neutral	neutral
3	suddenly, something amazing happened	surprise	surprise
4	we dare not obey your orders.	fear	fear
5	the place burnt like fire, and the poison entered into his blood	sadness	fear

As we noticed in sentence 1 and sentence 5, the model struggles to understand such ambiguous and complex sentences. In these two examples, sentence 1 is likely to have been predicted as happiness due to the presence of phrases/words like 'beautiful maiden' while sentence 5 could go either way (fear or sadness) even if it was left to human judgement.

C. Discussion

Multi-task BERT. Other than the aforementioned experiments, we also experimented on a multi-task BERT, which is motivated by Liu et al's work [21]. In our very own context, the Multi-task BERT learns a model to simultaneously predict the dimensional and categorical emotions at the same time (i.e., four outputs: Valence, Arousal, Dominance, and Category). This advantage presents some savings in terms of the number of parameters learned

and the inference time as well. Results in Table V show that the Multi-task BERT is capable of learning features from the sentences of stories reasonably well with an F1-score (EmoTales) and MAE/MSE (120 Stories) that matches the best performing single-task models. We expect better performances if the contributions of the MSE and categorical cross-entropy losses are properly balanced.

TABLE V
EXPERIMENT RESULT FOR EMOTION RECOGNITION FOR CATEGORY EMOTION AND DIMENSIONAL EMOTION

Model		EmoTales			120 Stories=(Inference Only)	
		8 Category	VAD		VAD	
Encoder	Decoder	F1-score	MAE	MSE	MAE	MSE
MLP		0.0719	0.1000	0.0164	0.1287	0.1287
Attention	FC	0.4424	0.1114	0.0206	0.1403	0.0409
Transformer	LSTM	0.4208	0.0966	0.0154	0.1113	0.0268
BERT	FC	0.5323	0.0887	0.0136	0.0961	0.0164
	LSTM	0.5467	0.0911	0.0136	0.1237	0.1237
	Bi-LSTM	0.5287	0.0846	0.0122	0.1228	0.0286
RoBERTa	FC	0.4892	0.0912	0.0143	0.1007	0.0160
	LSTM	0.4928	0.0967	0.0150	0.1291	0.0342
	Bi-LSTM	0.5035	0.0982	0.0142	0.1267	0.0300
ALBERT	FC	0.4424	0.1020	0.0175	0.1163	0.0207
	LSTM	0.4424	0.0960	0.0151	0.1276	0.0308
	Bi-LSTM	0.4424	0.0955	0.0146	0.1275	0.0297
ELMo	FC	0.4064	0.1025	0.0176	0.1167	0.0198
	LSTM	0.3844	0.0958	0.0166	0.3326	0.0249
	Bi-LSTM	0.4244	0.0984	0.0160	0.1688	0.0272
Multitask-BERT	FC	0.4856	0.0957	0.0147	0.1304	0.0215
	LSTM	0.5179	0.0870	0.0126	0.1114	0.0170
	Bi-LSTM	0.5431	0.0974	0.0152	0.1330	0.0225
EmoStory* (Cat → VAD)		-	0.0917	0.0142	0.1053	0.0174
EmoStory* (VAD → Cat)		0.4460	-	-	-	-

IV. CONCLUSION

Computational analysis of emotions in narrative stories is an understudied research domain with great potential in future applications. This paper presents several related problems that can be comprehensively integrated. Firstly, we propose an encoder-decoder framework to predict sentence-level emotion in stories in dimensional and categorical spaces, achieving reasonable accuracy. We also introduce a multi-head self-attention model, which can translate emotional representation from one space to the other and vice versa to a good measure of success.

The preliminary idea of EmoStory, a seamless prediction of both dimensional and categorical states of the emotion of stories, is trialed in this work, outlining a potential avenue of research in narrative text. We envision that in the future, such techniques could be extendable to model the personality or emotional state of characters in stories, which could benefit the affective assessment of experiences as well as the creation of emotive avatars and virtual worlds.

REFERENCES

[1] R. W. Picard, *Affective computing*. MIT press, 2000.

[2] T. Pólya and I. Csertő, "Emotion Recognition Based on the Structure of Narratives," *Electronics (Basel)*, vol. 12, no. 4, p. 919, 2023, doi: 10.3390/electronics12040919.

[3] S. Buechel and U. Hahn, "Emotion analysis as a regression problem–dimensional models and their implications on emotion representation and metrical evaluation," in *ECAI 2016*, IOS Press, 2016, pp. 1114–1122. doi: 10.3233/978-1-61499-672-9-1114.

[4] S. Ghosh, A. Ekbal, and P. Bhattacharyya, "VAD-assisted multi-task transformer framework for emotion recognition and intensity prediction on suicide notes," *Inf Process Manag*, vol. 60, no. 2, p. 103234, 2023, doi: 10.1016/j.ipm.2022.103234.

[5] C. Liu, M. Osama, and A. De Andrade, "DENS: A dataset for multi-class emotion analysis," *arXiv preprint arXiv:1910.11769*, 2019, doi: 10.48550/arXiv.1910.11769.

[6] E. Tromp and M. Pechenizkiy, "Rule-based emotion detection on social media: putting tweets on Plutchik's wheel," *arXiv preprint arXiv:1412.4682*, 2014, doi: 10.48550/arXiv.1412.4682.

[7] M. Wyczesany and T. S. Ligeza, "Towards a constructionist approach to emotions: verification of the three-dimensional model of affect with EEG-independent component analysis," *Exp Brain Res*, vol. 233, pp. 723–733, 2015, doi: 10.1007/s00221-014-4149-9.

[8] C. Strapparava and R. Mihalcea, "Learning to identify emotions in text," in *Proceedings of the 2008 ACM symposium on Applied computing*, 2008, pp. 1556–1560. doi: 10.1145/1363686.1364052.

[9] A. Neviarouskaya, H. Prendinger, and M. Ishizuka, "Compositionality principle in recognition of fine-grained emotions from text," in *Proceedings of the International AAAI Conference on Web and Social Media*, 2009, pp. 278–281. doi: 10.1609/icwsm.v3i1.13987.

[10] V. Francisco, R. Hervás, F. Peinado, and P. Gervás, "EmoTales: creating a corpus of folk tales with emotional annotations," *Lang Resour Eval*, vol. 46, pp. 341–381, 2012, doi: 10.1007/s10579-011-9140-5.

[11] Y. Lim, K.-W. Ng, P. Naveen, and S.-C. Haw, "Emotion Recognition by Facial Expression and Voice: Review and Analysis," *Journal of Informatics and Web Engineering*, vol. 1, no. 2, pp. 45–54, 2022, doi: 10.33093/jiwe.2022.1.2.4.

[12] A. Mehrabian and J. A. Russell, *An approach to environmental psychology*. the MIT Press, 1974.

[13] S. Buechel and U. Hahn, "Emotion representation mapping for automatic lexicon construction (mostly) performs on human level," *arXiv preprint arXiv:1806.08890*, 2018, doi: 10.48550/arXiv.1806.08890.

[14] S. Park, J. Kim, S. Ye, J. Jeon, H. Y. Park, and A. Oh, "Dimensional emotion detection from categorical emotion," *arXiv preprint arXiv:1911.02499*, 2019, doi: 10.48550/arXiv.1911.02499.

[15] M. Abdul-Mageed and L. Ungar, "Emonet: Fine-grained emotion detection with gated recurrent neural networks," in *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, 2017, pp. 718–728. doi: 10.18653/v1/P17-1067.

[16] Matthew E. Peters *et al.*, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter*

of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, Jun. 2018, pp. 2227–2237.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018, doi: 10.48550/arXiv.1810.04805.

[18] A. Vaswani *et al.*, "Attention is all you need," *Adv Neural Inf Process Syst*, vol. 30, 2017.

[19] S. Zhu, S. Li, and G. Zhou, "Adversarial attention modeling for multi-dimensional emotion regression," in *Proceedings of the 57th annual meeting of the association for computational linguistics*, 2019, pp. 471–480. doi: 10.18653/v1/P19-1045.

[20] Y. Yang, D. Zhou, Y. He, and M. Zhang, "Interpretable relevant emotion ranking with event-driven attention," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, 2019, pp. 177–187. doi: 10.18653/v1/D19-1017.

[21] Y. Liu *et al.*, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019, doi: 10.48550/arXiv.1907.11692.

[22] Z. Wu, X. Zhang, T. Zhi-Xuan, J. Zaki, and D. C. Ong, "Attending to emotional narratives," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, IEEE, 2019, pp. 648–654. doi: 10.1109/ACII.2019.8925497.

[23] D. Cortiz, "Exploring transformers in emotion recognition: a comparison of bert, distillbert, roberta, xlnet and electra," *arXiv preprint arXiv:2104.02041*, 2021, doi: 10.48550/arXiv.2104.02041.

[24] V. Francisco, P. Gervás, and F. Peinado, "Ontological reasoning to configure emotional voice synthesis," in *Web Reasoning and Rule Systems: First International Conference, RR 2007, Innsbruck, Austria, June 7-8, 2007. Proceedings 1*, Springer, 2007, pp. 88–102. doi: 10.1007/978-3-540-72982-2_7.

[25] M. M. Bradley and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *J Behav Ther Exp Psychiatry*, vol. 25, no. 1, pp. 49–59, 1994, doi: 10.1016/0005-7916(94)90063-9.

[26] James Carney and Cole Robertson, "4000 stories with sentiment analysis dataset." Accessed: May 19, 2023. [Online]. Available: https://brunel.figshare.com/articles/dataset/4000_stories_with_sentiment_analysis_dataset/7712540

[27] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015, doi: 10.48550/arXiv.1508.04025.

[28] X. Liu, P. He, W. Chen, and J. Gao, "Multi-task deep neural networks for natural language understanding," *arXiv preprint arXiv:1901.11504*, 2019, doi: 10.48550/arXiv.1901.11504.

[29] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019, doi: 10.48550/arXiv.1909.11942.

[30] J. Wei and K. Zou, "Eda: Easy data augmentation techniques for boosting performance on text classification tasks," *arXiv preprint arXiv:1901.11196*, 2019, doi: 10.48550/arXiv.1901.11196.