EmoStory: Emotion Prediction and Mapping in Narrative Stories

Seng-Wei Too\textsuperscript{a}, John See\textsuperscript{b}, Albert Quek\textsuperscript{c}, Hui-Ngo Goh\textsuperscript{c,}\textsuperscript{*}

\textsuperscript{a} Auronex Sdn Bhd, Bandar Bukit Jalil, 57000, Kuala Lumpur, Malaysia
\textsuperscript{b} School of Mathematical and Computer Sciences, Heriot-Watt University Malaysia, Putrajaya, Malaysia
\textsuperscript{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.


International Journal on Informatics Visualization is licensed under a Creative Commons Attribution-Share Alike 4.0 International License.

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.
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 axis 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 mechanisms—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 recursion 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 Yc∈C = {affection, anger, bravery, fear, happiness, sadness, surprise, neutral}; and Ye = [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:

\[ Y_c = f_c(X) \]  (1)
\[ Y_e = f_{v,a,d}(X) \]  (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 Yc category and vice versa. The problem is summarized as follows:

\[ Y_c = g((v, a, d)) \]  (3)
\[ Y_e = \tilde{g}(c) \]  (4)

where g and \( \tilde{g} \) 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 schema (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(·), takes an input e with three dimensions (v, a, d) and propagates it through h number of self-attention heads and fully-connected layer. At the output layer, a softmax classifier S_{softmax}() is applied to predict the probability \( P_c \) of each emotion \( c \) and the final mapped category \( Y_c \) is obtained by:
\[ \hat{Y}_c = S_{cm}(e) = y_c \in \mathcal{P}_c \]  

(5)

where

\[ \mathcal{P}_c = \frac{\exp(w^T(e_j))}{\sum \exp(w^T(e_j))} \]  

(6)

and \( W \) are parameters learned from model \( M \). Fig. 2 shows the general flow of the Multi-head Attention Network employed for 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 \( E \) 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, \( \hat{f}(X) \), 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 \( Y_e \).

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, \( f_{emd} \). 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, \( Y_c \):

\[ Y_c = g(f_e(X)) \]  

(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. Experiments 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 models: (i) FC: A basic single fully connected layer of 128 values, and queries. Dot-product self-attention is then applied, 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.

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 EMOTALE.

<table>
<thead>
<tr>
<th>Model</th>
<th>EmoTales</th>
<th>ERMDB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$c \rightarrow c$</td>
<td>$c \rightarrow e$</td>
</tr>
<tr>
<td>F1-Score</td>
<td>MAE</td>
<td>pmono</td>
</tr>
<tr>
<td>MLP [6]</td>
<td>0.37</td>
<td>0.1003</td>
</tr>
<tr>
<td>MLP + 1H-SelfAtt</td>
<td>0.59</td>
<td>0.0856</td>
</tr>
<tr>
<td>MLP + 2H-SelfAtt</td>
<td>0.53</td>
<td>0.0870</td>
</tr>
<tr>
<td>MLP + 4H-SelfAtt</td>
<td>0.59</td>
<td>0.0845</td>
</tr>
<tr>
<td>MLP + 8H-SelfAtt</td>
<td>0.51</td>
<td>0.0847</td>
</tr>
</tbody>
</table>

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 unidirectional 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 $5 \times 10^{-4}$) 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 experimental 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 $\rho$ 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 $7 \times 10^{-4}$, 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
ERMA baseline comparison in monolingual setting

<table>
<thead>
<tr>
<th>Model</th>
<th>V</th>
<th>A</th>
<th>D</th>
<th>Joy</th>
<th>Ang</th>
<th>Sad</th>
<th>Fear</th>
<th>Dis</th>
<th>AvgVAD</th>
<th>AvgBE5</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline-Reported</td>
<td>0.956</td>
<td>0.754</td>
<td>0.810</td>
<td>0.932</td>
<td>0.855</td>
<td>0.816</td>
<td>0.838</td>
<td>0.752</td>
<td>0.8400</td>
<td>0.8386</td>
<td>1.6786</td>
</tr>
<tr>
<td>f Baseline-Reproduced</td>
<td>0.9558</td>
<td>0.7545</td>
<td>0.8104</td>
<td>0.9323</td>
<td>0.8555</td>
<td>0.8155</td>
<td>0.8377</td>
<td>0.7516</td>
<td>0.8402</td>
<td>0.8385</td>
<td>1.6787</td>
</tr>
<tr>
<td>MLP + 1H-SelfAtt</td>
<td>0.9539</td>
<td>0.7395</td>
<td>0.8002</td>
<td>0.9109</td>
<td>0.8507</td>
<td>0.8116</td>
<td>0.8346</td>
<td>0.7459</td>
<td>0.8312</td>
<td>0.8307</td>
<td>1.6619</td>
</tr>
<tr>
<td>MLP + 2H-SelfAtt</td>
<td>0.9487</td>
<td>0.7218</td>
<td>0.7745</td>
<td>0.9204</td>
<td>0.842</td>
<td>0.7985</td>
<td>0.8241</td>
<td>0.7291</td>
<td>0.8150</td>
<td>0.8228</td>
<td>1.6378</td>
</tr>
<tr>
<td>MLP + 4H-SelfAtt</td>
<td>0.9486</td>
<td>0.7205</td>
<td>0.7802</td>
<td>0.9199</td>
<td>0.8423</td>
<td>0.7988</td>
<td>0.8247</td>
<td>0.7261</td>
<td>0.8164</td>
<td>0.8223</td>
<td>1.6388</td>
</tr>
<tr>
<td>MLP + 8H-SelfAtt</td>
<td>0.9487</td>
<td>0.7143</td>
<td>0.7775</td>
<td>0.9201</td>
<td>0.8403</td>
<td>0.7944</td>
<td>0.8216</td>
<td>0.7216</td>
<td>0.8135</td>
<td>0.8196</td>
<td>1.6331</td>
</tr>
</tbody>
</table>

TABLE III
ERMA baseline comparison in cross-lingual setting

<table>
<thead>
<tr>
<th>Model</th>
<th>V</th>
<th>A</th>
<th>D</th>
<th>Joy</th>
<th>Ang</th>
<th>Sad</th>
<th>Fear</th>
<th>Dis</th>
<th>VAD</th>
<th>BE5</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline-Reported</td>
<td>0.9510</td>
<td>0.6740</td>
<td>0.9260</td>
<td>0.829</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>f Baseline-Reproduced</td>
<td>0.9506</td>
<td>0.6586</td>
<td>0.9019</td>
<td>0.8264</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLP + 1H-SelfAtt</td>
<td>0.9485</td>
<td>0.6731</td>
<td>0.9007</td>
<td>0.8258</td>
<td>0.7743</td>
<td>0.7932</td>
<td>0.7080</td>
<td></td>
<td>0.8108</td>
<td></td>
<td>1.6112</td>
</tr>
<tr>
<td>MLP + 2H-SelfAtt</td>
<td>0.9433</td>
<td>0.6382</td>
<td>0.9160</td>
<td>0.8373</td>
<td>0.7920</td>
<td>0.8042</td>
<td>0.7265</td>
<td>0.7908</td>
<td>0.8152</td>
<td></td>
<td>1.6060</td>
</tr>
<tr>
<td>MLP + 4H-SelfAtt</td>
<td>0.9426</td>
<td>0.6645</td>
<td>0.9160</td>
<td>0.8365</td>
<td>0.7842</td>
<td>0.7992</td>
<td>0.7192</td>
<td>0.8036</td>
<td>0.8110</td>
<td></td>
<td>1.6146</td>
</tr>
<tr>
<td>MLP + 8H-SelfAtt</td>
<td>0.9407</td>
<td>0.6580</td>
<td>0.9130</td>
<td>0.8376</td>
<td>0.7858</td>
<td>0.8033</td>
<td>0.7240</td>
<td>0.7994</td>
<td>0.8127</td>
<td></td>
<td>1.6121</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th># Test Sample</th>
<th>True</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 she gave him the mirror in his hand, and he saw there in the likeness of the most beautiful maiden on earth mounts, i too will try my fortune,</td>
<td>surprise</td>
<td>happiness</td>
</tr>
<tr>
<td>2 if that is the ladder by which one 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, 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.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.
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


